Automating three-dimensional osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography

Santeri Rytky, santeri.rytky@oulu.fi
Pro gradu thesis, 35 ECTS
Physics Degree Program
Faculty of Science, University of Oulu
7.5.2019
Supervisors: Simo Saarakkala,
Aleksei Tiulpin

UNIVERSITY OF OULU

Table of Contents

Abstract

Osteoarthritis (OA) is a joint disease affecting hundreds of millions of people worldwide. In basic research, accurate *ex vivo* measures are needed for assessing OA severity. The standard method for this is the histopathological grading of stained thin tissue sections. However, the methods are destructive, time-consuming, do not describe the full sample volume and provide subjective results. Contrast-enhanced micro-computed tomography (CEμCT) –based grading with phosphotungstic acid -stain was previously developed to address some of these issues. Aim of this study was to investigate the possibility of automating this process.

Osteochondral tissue cores were harvested from total knee arthroplasty patients (n = 34, $N_{patients}$ = 19, Ø = 2 mm, n = 15, $N_{patients}$ = 5, Ø = 4 mm) and asymptomatic cadavers (n = 30, $N_{patients}$ = 2, Ø = 4 mm). Samples were imaged with CEμCT, reconstructed and graded manually. Subsequently, the reconstructions were loaded into an *ad hoc* developed Python software, where volumes-of-interest (VOI) were extracted from different cartilage zones: surface zone (SZ), deep zone (DZ) and calcified zone (CZ) and collapsed into two-dimensional texture images.

Normalized images underwent Median Robust Extended Local Binary Pattern (MRELBP) -algorithm to extract the features, with subsequent dimensionality reduction. Ridge and logistic regression models were trained with L2 regularization against the ground truth for the small samples (Ø = 2 mm) using leave-one-patient-out cross-validation. Trained models were then evaluated on the large samples (Ø = 4 mm). Performance of the models were assessed using Spearman's correlation, Area under the Receiver Operating Characteristic Curve (AUC) and Average Precision (AP).

Highest performance on both models was for the SZ. Strong correlation was observed on ridge regression ($\rho = 0.68$, $p < 0.0001$), as well as high AUC and AP values for the logistic regression (AUC = 0.92, AP = 0.89) for the small samples. Using the large samples, similar findings were observed with slightly reduced values ($\rho = 0.55$, $p = 0.0001$, AUC = 0.86, AP = 0.89). Moderate results were observed for CZ and DZ models ($\rho = 0.54$ and 0.38, AUC = 0.77 and 0.72, AP = 0.71 and 0.50, respectively). Evaluation on the large samples resulted in performance decrease on CZ models ($\rho = 0.29$, AUC = 0.63, AP = 0.62), while surprisingly performance increased on DZ logistic regression model ($\rho = 0.34$, AUC = 0.72, AP = 0.83).

Obtained results indicate that automating the 3D CEμCT histopathological grading is feasible. However, with low number of samples, models are better suited for binary detection of

sample degenerative features, rather than predicting a detailed grade. To facilitate model generalization on new data, similar data acquisition protocol should be used on all samples. The proposed methods have potential to aid OA researchers and pathologists in 3D histopathological grading, introducing more objectivity to the grading process. This thesis presents the conducted study in detail, and provides an extensive review related to the osteochondral unit, CEμCT imaging, as well as statistical learning machines.

Keywords: *osteoarthritis, histopathological grading, contrast-enhanced micro-computed tomography, machine learning, cartilage, textural analysis*

Author's contribution

The publication related to this thesis is an original research paper on contrast-enhanced micro-computed tomography of articular cartilage [1]. The author has contributed in development of the presented methods used to automate the 3D grading of articular cartilage. The author was the main writer of this thesis. During development of the prototype software referred to in this thesis, the author was responsible for implementing the developed automatic 3D grading functionalities.

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
AC	Articular cartilage
AP	Average precision
AUC	Area under receiver operating characteristic curve
CA4+	Cationic contrast agent
CEμCT	Contrast-enhanced micro-computed tomography
CCD	Charge-coupled device
CNN	Convolutional neural network
CPU	Central processing unit
CT	Computed tomography
CZ	Calcified zone
DZ	Deep zone
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacetic acid
LASSO	Least absolute shrinkage and selection operator
LOO	Leave-one-out
LOPO	Leave-one-patient-out
LS	Least squares
μCT	Micro-computed tomography
MRELBP	Median robust extended local binary pattern
MSE	Mean squared error
OA	Osteoarthritis
OARSI	Osteoarthritis Research Society International
PBS	Phosphate buffered saline
PCA	Principal component analysis
PCM	Pericellular matrix
PG	Proteoglycan
PRC	Precision-recall curve
PSNR	Peak signal-to-noise ratio
PTA	Phosphotungstic acid
riu2	Rotation invariant uniform mapping
ROC	Receiver operating characteristic
SCB	Subchondral bone
SSIM	Structural similarity index
SVD	Singular value decomposition
SZ	Superficial zone / Surface zone
TKA	Total knee arthroplasty
TPE	Tree of Parzen estimators

1. Introduction

Osteoarthritis (OA) is a joint disease causing major economic burden and affecting millions of people globally [2]. It is hypothesized that OA is a group of diseases with different phenotypes [3] and early detection of the disease is important to be able to stop its progression [4]. Basic OA research can help to understand the disease characteristics, as well as aid in drug and biomarker development. This requires accurate *ex vivo* assessment of the early degenerative changes of cartilage, and it can be conducted by optical imaging of histochemically stained tissue sections. The current gold standard method is the OARSI grading system [5], developed specifically for histochemical imaging and allowing early assessment of OA.

Histological methods have multiple limitations. They require destruction of the sample and the process of obtaining stained sections can take weeks. Furthermore, thin sections often do not represent the full volume of the sample, and assessing the grades requires multiple trained professionals to achieve a result with sufficiently reduced user bias. Nieminen et al [6] have previously developed a nondestructive contrast-enhanced micro-computed tomography (CEμCT) grading system with a collagen-specific contrast agent to allow volumetric grading of osteochondral tissue. Automating this process could introduce more objective grading results.

Machine learning methods have been used in OA research in clinical environment, to estimate disease severity [7] and prognosis [8,9]. However, implementation of machine learning for basic OA research has received less attention. The present study is focused on this area. Previously, quantitative 3D assessment of osteochondral tissues have been proposed for articular cartilage (AC) surface [10,11], calcified cartilage [12] and full cartilage depth [13]. However, the studies utilize only internal testing [13] or are limited to assessing a single cartilage zone [10-12].

Aim of this thesis is to present a method that helps automate the CEμCT grading process using machine learning. Further, the generalization of this method on unseen data is assessed using an independent Test set. I hypothesize that the degenerative features of the selected osteochondral zones can be captured using texture analysis and moreover, utilized in statistical learning algorithms to predict the degeneration of new samples.

2. Osteochondral tissue and osteoarthritis

2.1 Articular cartilage

Articular cartilage (AC) is a connective tissue at the ends of long bones, consisting of 2-4mm thick layer of non-calcified hyaline cartilage [14]. It allows almost frictionless movement in articulating joints such as knee, hip and hand. It does not include any blood vessels, neurons or lymph nodes, and nutrient delivery is highly relying on diffusion from synovial fluid assisted by mechanical loading of the joint [15]. Lack of subchondral blood vessels also leads to limitations in cartilage healing capabilities [14].

Main components of AC are water, collagen and proteoglycans (PG). Glycosaminoglycan side chains of PGs are negatively charged. Small amounts of other proteins are also present. Collagen fibers (mainly type II) are bundled [16] and form the macrostructure of AC with different fiber orientations depending on location. Proteoglycans (PG, mainly aggrecan) are negatively charged glycoproteins bound to collagen fibers. High electric charge of PGs binds to water and contributes to viscoelastic properties of AC. Together, collagen fiber network and PGs form the extracellular matrix (ECM) of AC. [14]

AC is divided in zones with different functional and compositional properties (Figure 1) [14,16]. Superficial zone (SZ) extends from cartilage surface roughly 10-20% of the AC depth. On SZ, collagen fibers are oriented parallel to cartilage surface. Below SZ is the transitional zone, 40-60% of cartilage depth. In this zone, fibers are randomly oriented. Lowest non-calcified zone is the deep zone (DZ), extending 30-40% of cartilage depth. Zonal variations of AC have been known for a long time. Benninghoff [17] was the first to suggest an arcade-formation of long collagen fibers along the cartilage depth and confirming observations were made early [16,18].

Immediately below DZ, a tidemark marks the border of calcified cartilage zone (CZ). This zone consists of mineralized AC and acts as important connector between AC and subchondral bone (SCB). Collagen fibrils (type II) are anchored on the CZ to the SCB layer (fibers do not extend to the SCB plate) [19]. Type X collagen is also present at CZ. Border between SCB and CZ is known as cement line [15].

Chondrocytes are specialized cartilage cells that are responsible of maintaining the ECM. They take up to 2% of the AC volume [20]. They are shaped flat in the SZ and are larger and more spherical in deeper zones of AC. Chondrocytes of DZ are arranged in columns perpendicular to AC surface [15]. Chondrocytes generate new PGs and have generated a surrounding matrix known as the pericellular matrix (PCM). Chondron is a unit formed by chondrocyte and its PCM, introduced also by Benninghoff [17]. Chondrocytes have limited capability of replicating, which diminishes the cartilage healing abilities [14]. Chondrocytes are active cells responding to biochemical and mechanical stimuli and possess variety of ion channels that interact with chondrocyte environment [21].

2.2 Subchondral bone

SCB main anatomical components include SCB plate, and spongy trabecular bone. SCB plate is often defined as the bony layer separating AC from porous trabecular bone, while other definitions also exist depending on source. CZ is separated from SCB plate by the cement line. Plate thickness varies depending of the joint. Below the SCB plate, resides trabecular bone layer consisting of porous structures. Both of these bone-made structures consists of type I collagen. [19]

Contrary to AC, SCB tissues include both blood vessels and nerves. Vascular channels extending through SCB from bone marrow all the way to CZ connect deep cartilage layers to marrow. These channels can provide nutrients to calcified cartilage layer, while others extend only to bone layer [22]. There is spatial variation in number of channels and regions of high load, e.g. weight-bearing area on medial tibial plateau has an increased number. Areas without these vascular channels have the synovial fluid as the only source of nutrition. [19]

SCB has a major role in withstanding mechanical load in the joint. Healthy SCB attenuates 30% of joint load compared to 1-3% of AC [19]. Other functions of SCB include AC support and nutrition (deep layers). Osteochondral structure is further illustrated in Figure 1.

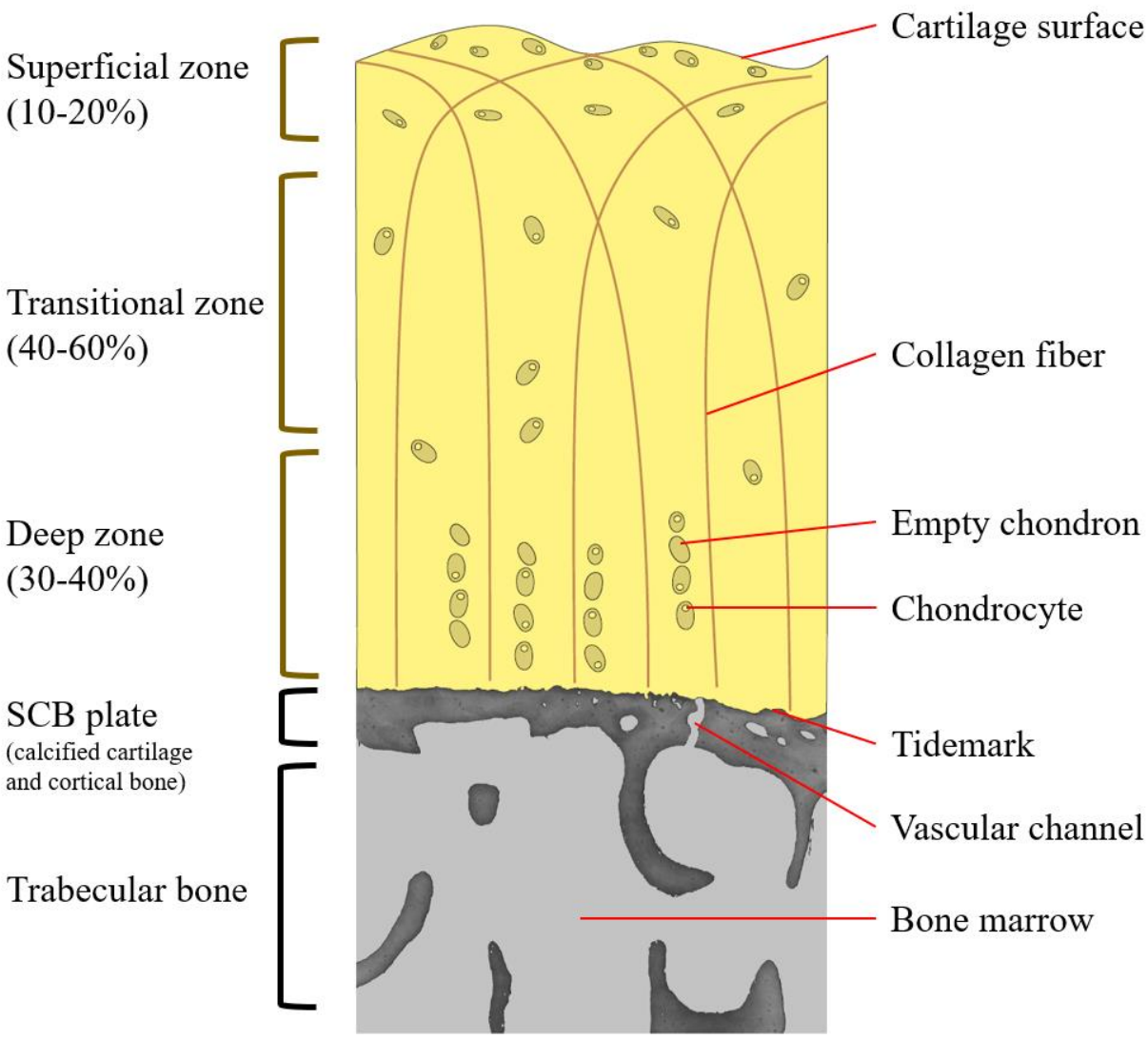

Figure 1. Illustration of AC and SCB composition and structure. Different osteochondral zones and features are described.

2.3 Osteoarthritis

Osteoarthritis (OA) is a joint disease, involving the full joint on organ level, including patient symptoms, structural remodeling and tissue inflammation [23,24]. It is the most common cause for disabilities affecting 12.3% (France, 2008) to 21.6% (United States, 2003-2005) of population [2]. At the very end stage, knee OA is treated using total knee arthroplasty (TKA), replacing the joint with an implant.

In France (2002), direct expenses caused by OA were 1.6 billion, while the annual cost of hip and knee total replacement in United States (2004) was 7.9 billion. Individual risk factors of

OA include age, gender (women have higher risk), obesity, genetics and diet. Joint injuries and abnormal loading, for example due to occupation are also contributing to OA risk. [2]

OA has long thought to be only a cartilage-related disease. It has been recognized that there is a strong crosstalk between SCB and AC [25]. Changes in AC due to OA progression include surface fibrillation, chondrocyte clustering and PG loss on superficial zone on the early stages, while fissures and cartilage ECM loss start occurring on later stages of disease [5]. Changes in bone include sclerosis and remodeling due to cell activation [26]. This results in SCB plate thickening as well as calcification of deep AC layers. Local tissue strains increase during progression of OA [27].

2.4 Histopathological grading

Histopathological grading is used for assessing OA severity *ex vivo*. To allow the grading of osteochondral tissues, standard method is to use histochemical staining. First, samples are fixated to prevent sample degradation. Most commonly used fixative is neutral buffered formalin stabilizing amino acids and preserving tissue structure [28]. It has been shown to preserve PG content of cartilage tissues well [29]. To prevent chondrocyte shrinkage in cartilage studies, fixative osmolarity should be set according to extracellular fluid (approximately 280 mOsm). This can be achieved using ruthenium hexamine trichloride solution with cacodylate buffer [30].

After osteochondral samples are fixed, decalcification of tissues is often performed using ethylenediaminetetraacetic acid (EDTA). Other solutions such as formic acid also exist. Decalcified samples are then washed and dehydrated in ascending ethanol series. This is followed by paraffin embedding to allow thin sectioning by microtome. Cryosections on fresh samples can also be used on joint histology. Sectioning of samples is performed with a microtome and can be focused on areas of interest or alternatively, sections of the full sample can be made. Slice thickness varies usually from 3 to 6 µm. Tissue sections are placed on microscope slides and stained using preferred solution. Safranin-O and Toluidine blue are PG-specific stains, while e.g. Picrosirius Red can be used to assess collagen distribution [6]. [31]

There are multiple grading systems established for evaluating OA progression, most commonly used being OARSI grade [5], Histologic Histochemical Grading System (Mankin score) [32] and O'Driscoll score [33]. They are based on assessing histochemically stained two-dimensional (2D) thin sections.

Mankin score was developed based on late stage OA features, and it is insensitive to the early changes of OA [34]. There are several modified Mankin scores trying to address this problem. OARSI grading and staging system was created to allow more accurate assessment of early stages of OA and make evaluation less subjective [5]. Early signs of OA detected with OARSI are significantly related to differences in cartilage biomechanical properties [35]. This system is heavily focused on AC degeneration, and SCB remodeling is fundamentally linked in this process [4]. OARSI grade has currently been recognized as gold standard for *ex* vivo assessment of OA progression.

While histological methods create high-resolution microscopic images providing chemical and structural information of the sample, they possess many limitations. They require sample destruction, are time-consuming [34], do not capture three-dimensional (3D) properties of the full sample volume, and the final evaluation of grades is subjective without assessment of multiple trained graders. To address these issues, Nieminen et al [6] developed recently a grading system based on 3D evaluation of osteochondral samples using contrast enhanced micro-computed tomography (CEμCT). This thesis studies the possibilities of automating the developed grading system.

3. Micro-computed tomography

3.1 The history of X-rays

Wilhelm Röntgen discovered the X-rays already in 1895. During the next year, technology was already used for imaging of human anatomy (Hoffmans and van Kleef). Ever since, X-rays have been used extensively in medical imaging applications. First computed tomography (CT) systems were developed in 1970s by Sir Geoffrey Hounsfield to allow volumetric X-ray imaging. At the time of first generation X-ray devices, long exposure times were required to obtain images with sufficient quality. With modern systems, exposure times have reduced from 90 minutes to 21 milliseconds, effectively reducing the radiation dose from 74 mGy to 0.05mGy [36]. On the other hand, image quality has also substantially increased (Figure 2). In the present day, X-ray imaging techniques can safely provide substantial amount of structural information from different tissues.

Figure 2. Comparison of first generation X-ray imaging technology to its modern counterpart. Sharper image can be acquired with a fraction of the radiation dose using modern X-ray systems. Image on the left was acquired by reproducing experiments of Hoffmans and van Kleef in 1896. Right side shows an image from a modern X-ray source. Both images were taken using a modern radiographic plate. Reprinted from Kemerink et al. [36] with permission from Radiological Society of North America.

3.2 X-ray physics

X-rays, as well as any other form of electromagnetic radiation consist of photon quanta. Concept of the quantum was first described by Albert Einstein in 1905. Photon wavelength λ and frequency v are related as in the wave equation $c = \lambda v$, where c is the speed of light. Quantum energy is given by $E = hv$, where h denotes the Planck constant. Photon energies used in medical imaging range roughly from tens of kilovolts to 150 keV [37].

In medical X-ray imaging, fundamental interactions between photons and tissue include photoelectric effect, Compton and Thomson (a.k.a. Rayleigh) scattering. Overall contrast in X-ray images is created by absorption of photons through photoelectric effect. If photon energy exceeds the binding energy of inner shell (often K-shell) electron in the tissue atom, photon is absorbed and electron is released. The gap in the atom's electron structure is filled by outer shell electron, and either characteristic X-ray photon or Auger electron (another outer shell electron) is released. However, the absorption of the incident photon is the phenomenon that mainly causes attenuation observed in the X-ray image. These absorption peaks can be seen as k-peaks in the X-ray absorption spectrum.

Compton (inelastic) scattering is another common photon interaction mechanism. In this case, incident photon collides with the tissue electron (or rarely with the atom nucleus), the photon loses some of its energy to the electron and scatters to another direction from the incident angle. Thomson (elastic) scattering is the low-energy limit for Compton scattering. In such case, no energy is lost to the recoil electron, however the scattering event causes direction of the photon to change. The scattering events cause the photon to land on an unexpected pixel on the X-ray detector and thus create noise on the measurement.

During X-ray imaging, sum of the events described above is measured on the X-ray detector. Attenuation on the imaged sample can be modeled using Beer-Lambert law or its derivatives

$$I = I_0 e^{-\mu x} = \int_0^{E_{max}} I_0(E) e^{-\int_0^d \mu(E,s)ds} dE \,. \tag{1}$$

In the original formula of the law, I_0 is the original X-ray intensity, I denotes the intensity after attenuation, μ = attenuation coefficient of the tissue and x is the sample thickness. In the more detailed version of the model (equation 1), it is taken into account that X-ray intensity often depends on the photon energy E (except for monochromatic sources, e.g. synchrotron facilities).

Also μ is depending on both the energy and spatial location on the sample, in case of heterogeneous samples (most biomedical samples are heterogeneous) according to $\mu = (\rho Z^4) / (AE^3)$ [38]. Here, ρ = tissue density, Z = atomic number and A = atomic mass. Thus, in X-ray imaging experiments, attenuation values measured on the detector can be modeled as line integrals along the incident X-ray path (equation 1).

3.3 Computed tomography imaging

CT imaging is a technique used to obtain volumetric information using X-ray imaging. It is based on taking multiple images of the subject incrementing the imaging angle. Usually images are taken on equal spacing for 180 or 360 degrees. In X-ray microtomography, or micro-computed tomography (μCT), measurement is conducted *ex vivo* for tissue specimen or for small animals *in vivo*.

Scanners used in μCT imaging consist of three main components: X-ray source, sample compartment and detector. In laboratory μCT setups, micro- or nanofocus X-ray tube is the most common type of source. X-rays are produced by accelerating electrons in the vacuum tube and colliding the electrons into a heavy metal anode (usually tungsten). This collision produces braking radiation (Bremsstrahlung) and sharp peaks known as characteristic X-rays, to the X-ray energy spectrum. Produced X-rays are then collimated using magnetic lenses to focus the X-rays on a small spot (1-10μm range, micro- and nanofocus setups). This allows controlled exposure of the sample (e.g. creating a point source in case of cone-beam geometry). Clinical CT devices use X-ray tubes with spots size around 300μm range and require low magnification in the imaging setup. [39]

In cone-beam geometry, when X-rays exit the collimator, they are assumed to distribute evenly in the sample chamber originating from a (very small) point source [40]. This creates an X-ray "cone", which illuminates the sample. Advantage of using cone-beam geometry is that the whole sample can be imaged using a single rotation. Clinical CT systems often use a fan-beam geometry, where line detector is used and images are taken incrementally in a two-dimensional setup. Modern clinical systems often use a spiral CT scan, invented by Willi Kalender [41]. In spiral CT, images are taken on a helical path around the patient. [39]

Geometric magnification is the most commonly used method to acquire high resolution details in μCT imaging. When the sample is moved closer to source, the "shadow" of the sample on the detector is enlarged. Some devices can move also the detector closer to sample, allowing better

preservation of spatial resolution. On *ex vivo* setups, sample is attached to a rotating holder, which allows changing the position of the sample and utilizing the geometrical magnification. Preclinical and clinical scanners are designed so that the source and detector rotate around the sample instead, mounted on a gantry [42].

After attenuation on the sample, X-rays are measured on the detector. Before the detector, a scintillator array converts the X-rays to visible light. Accuracy of the scintillator material can be described using a point-spread function [43]. Optionally, zooming optics can be used after the scintillator to provide additional magnification to light measured from the scintillator. Finally, the light is measured on a detector array. On the detector, absorption of light is converted to electric charge that is measured. Usually a charge-coupled device (CCD) camera is used as the detector. It provides high light sensitivity with expense of long readout times and high cost. Other detectors such as Complementary Metal Oxide Semiconductors or Active Matrix Flat Panel Imagers are sometimes used. [39]

3.4 Image reconstruction

To convert the obtained X-ray images into volumetric slices, the image data has to be reconstructed. Input for the reconstruction algorithm is a sinogram: combination of angular intensity profiles obtained from a detector line corresponding to reconstructed slice (Figure 3b). Most commonly used reconstruction algorithms are variants of Filtered back projection (FBP) developed by Feldkamp et al. [40]. FBP is based on Radon transform developed by Johann Radon in 1917, which is closely related to Fourier transform. [43]

The principle behind FBP is easily understood. Each intensity profile (Figure 3c) can be assumed to be evenly distributed on the X-ray path of the imaging setup (Figure 3a). When all imaged angles are taken into account, the intensities can be summed along the paths. With increasing number of projection, the sum starts to resemble a blurred version of the object in image space. To avoid the blurring artefact, absorption profiles should be filtered dampening low- and high-frequency components [43]. The low frequencies are related to the blurring artefact and high frequencies contribute to statistical noise. However, more advanced methods have been developed for geometric corrections in cone-beam imaging [44].

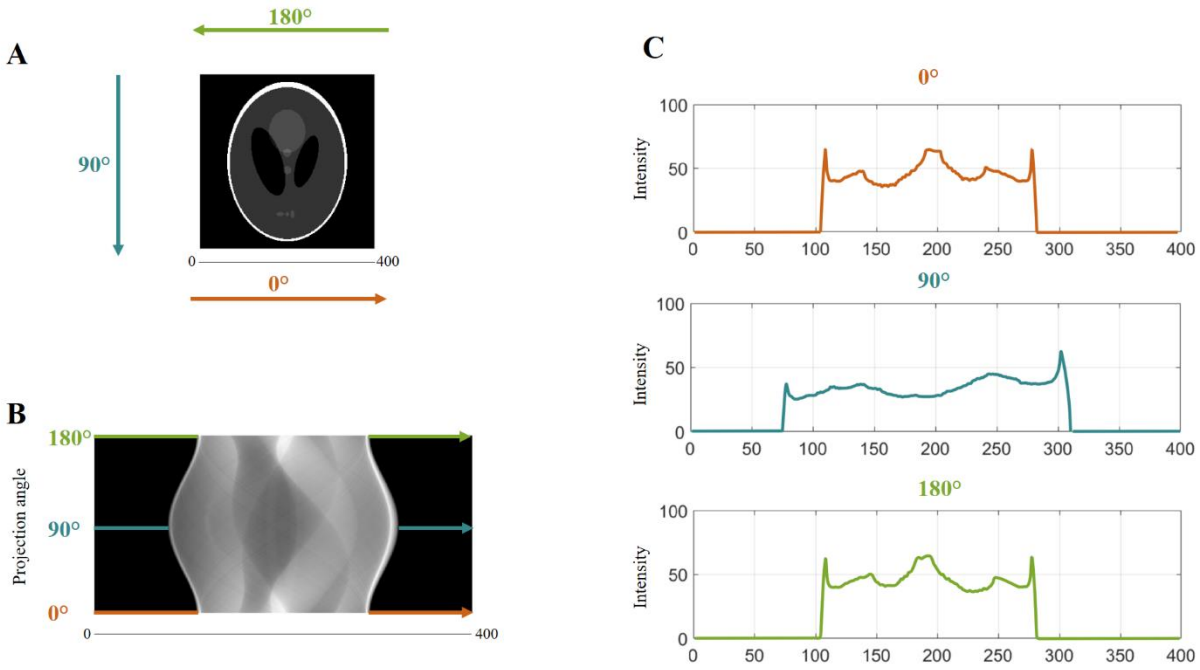

Figure 3. Illustration of creating sinogram from a 2D phantom image. Setup corresponds to image data acquired by a single detector row. X-ray images taken from a phantom **(A)** at different angles allow concatenating intensity histograms of individual angles into one image. Result is a sinogram where intensity profile is plotted against the imaging angle **(B)**. Intensity profiles from selected angles are shown **(C)**, profiles at 0^0 and 180^0 are mirror profiles from each other. Illustration was generated using Astra toolbox 1.8 (Matlab).

Alternative to reconstruction methods based on FBP, iterative reconstruction methods have been proposed to achieve accurate reconstructions from sparsely sampled CT data [45,46]. Iterative methods were used already in early years of CT imaging [47], but processing power of computers was limited and the algorithms did not get any practical use. Modern workstations have the computational capabilities required for IR techniques and it is possible to get reconstruction quality comparable to FBP algorithms with smaller number of projections. Development continues to further reduce the radiation dose and decrease the computational demand on higher resolution images.

In iterative reconstruction, instead of single reconstruction, multiple trials are calculated. Error metric is evaluated and sequential reconstructions are calculated until the algorithm converges to optimal solution. Physical models for phenomena such as statistical photon distribution or imaging geometry can be implemented in the algorithms. The iteration process consists of three main steps. Forward projection of the estimated slice is used to obtain a projection

image. This image is then compared against the original projection and correction term is estimated. The term is then used in calculating another back projection of the projection image to obtain new estimate of the reconstructed slice. Initial guess for the reconstruction can be obtained using an empty slice or FBP of the projection. [45]

Main limitation of the iterative reconstruction is related to the computational requirements. Thus, iterative methods may not be yet suitable for the most complicated CT imaging setups, e.g. multi-energy imaging. There are also problems in applying iterative methods in clinics. Radiologists are used in analyzing reconstructions obtained with FBP, and taking a new method in everyday use requires comprehensive studying of new or different features and artefacts in the reconstructions [45].

3.5 Contrast enhanced imaging

CT imaging is often thought only to be applicable for imaging calcified tissues such as bone. However, soft tissue contrast can be achieved using advanced techniques. Contrast agent-, multiple-energy- and phase-contrast imaging belong to these methods. Some of them are already in clinical use (iodine, barium, dual-energy imaging) [38]. Regarding μCT imaging, these methods are collectively referred as contrast-enhanced μCT (CEμCT).

Contrast agent CT imaging is based on measuring the distribution of the contrast agent on studied sample. Most contrast agent molecules consist of two parts: a heavy metal used to provide X-ray attenuation contrast and a tissue specific component that provides either biochemical or functional information of the studied tissue [38]. On clinical environment, toxicity of the contrast agent is also of concern.

There are multiple contrast agents developed that are suitable for imaging osteochondral tissue. Ioxaglate (Hexabrix™) is clinically approved and used to assess PG content of AC [48]. Ioxaglate content is inversely proportional to PGs, since ioxaglate is anionic and there is a repulsive interaction with PGs. *Ex vivo*, cationic contrast agent (CA4+) can be used for PG imaging [49]. It has advantage of having proportional stain, since it is cationic. However, there can be problems with contrast agent washout due to electric binding.

Phosphotungstic acid (PTA) is a histological stain that was first used as a CT contrast agent by Metscher et al [50]. It has been recently introduced as a collagen specific stain in AC [51,52] and a 3D grading system based on the imaging agent has been developed [6].

To separate multiple contrast agents from each other, multiple μCT scans can be conducted with different energies. Different scans should be planned so that X-ray energies are close to the k-edges of the contrast agents. On the other hand, contrast agents should be selected so that their k-edges are distinct from each other. For example, dual-energy or triple energy imaging can be used to separate two or three contrast agents [53,54]. Further, feasibility for dual-energy μCT on cartilage imaging without additional staining has been suggested [55].

Phase-contrast imaging is based on measuring refractive index decrement of the sample (real component of the refractive index, absorption imaging is related to the imaginary component) [43]. To measure the phase change, a coherent X-ray source is required. Phase contrast enhances different optical interfaces on the sample and is capable of providing contrast also in soft tissues, even on submicron scale [56]. Usually synchrotron facilities are required to achieve sufficient coherence of the X-rays for phase-contrast imaging, but intermediate results can be achieved even with laboratory sources [57].

To distinguish different material compositions from each other, phase retrieval algorithms have been developed for single distance [58] and multiple distance imaging [59]. Phase retrieval provides additional contrast to materials with different optical properties, e.g. allowing segmentation of different structures. Technique has been used for volumetric imaging of human vocal cords with comparable image quality to histology [56].

4. Machine learning

4.1 Machine learning in biomedical imaging

Statistical learning algorithms can be used in various biomedical problems such as predicting severity of a disease [60], prognosis (probability for progression) [8,61] or classification of different biomedical phenotypes [62,63]. Often a computationally intensive training process is required to learn the connections between extracted features and estimated outcomes. Once the training is completed, learned model can quickly be able to evaluate the process on new data, conducting an inference. Many issues have to be taken into account when developing the learning machine, related to overfitting, feature selection and performance evaluation. Data used in the model might be sparse or imbalanced (especially on biomedical applications), or not representative of the entire population studied.

Many biomedical questions are related to learning machines and most of them can be categorized into classification or regression. Goal in classification is to find patterns that separate the data into known classes with the classifier. In image analysis, this can mean segmenting a tumor from the medical image (e.g. X-ray, magnetic resonance image, histology) or defining areas of different tissue types. The classifier is trained using known labels of the classes, for example the malignant tumor areas can be segmented manually by a pathologists as the ground truth. This learning type, where correct answer is exactly known, is supervised learning [64]. Consequently, the model can evaluate when misclassifications are made or the distance from the ground truth. Most commonly used classifiers that utilize supervised learning are logistic regression, k-nearest neighbours, support vector machines and random forest. Unsupervised and semi-supervised learning types also exist, such as the clustering methods. K-means clustering is type of unsupervised learning used also in this thesis. [65]

Providing estimate of a given stage of a disease or a probability of a patient belonging to a specific class, are regression problems. Regression problems can be seen as analogous of a curve fitting problem. Linear regression and its variants are most famous of these type of learning machines. Most simplified classification problems, such as binary detection of healthy / diseased are often easier than quantifying exact disease stage or extent (similar regression problems can be reduced to classification using given threshold).

Regarding biomedical data, it should be noted that classes are usually not balanced. This is especially important when estimating errors of a given model. Let's say that we have a classification model that is trained to detect patients that have risk of disease progression. Data obtained consists of 100 patients with 97 healthy and 3 patients having risk of progression. If the model learns to label every patient as healthy, there is only 3% of misclassified patients. Clearly this is not the best way of evaluating performance in this case and more advanced techniques should be utilized. Another option is to oversample the patients with risk to balance the dataset, but this approach is either not without its caveats. [66]

4.2 Empirical risk minimization

To make predictions in supervised learning problems, an algorithm is needed to select the optimal model, which creates predictions $\hat{\boldsymbol{y}}$ from inputted features $\boldsymbol{x}$ based on ground truth $\boldsymbol{y}$. The difference of the predictions to the ground truth is estimated using a nonnegative loss function Φ. The risk (cost) associated in the model can be theoretically defined as the expectation E of Φ [67]:

$$R(\hat{y}) = E[\Phi(\hat{y}, y)] = \int \Phi(\hat{y}, y) dP(x, y)\,. \tag{2}$$

However, the expected loss (equation 2) cannot be solved analytically, since the probability distribution P of the features and ground truth is not known. Empirical risk R_{emp} can be estimated instead, by averaging the loss function over the samples used for training the model

$$R_{emp}(\hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \Phi(\hat{y}_i, y_i)\,. \tag{3}$$

Finally, the optimal solution can be found by choosing the model h which minimizes the empirical risk [65], described in equation (3)

$$h_{opt} = \arg\min \Phi(\hat{y}, y)\,. \tag{4}$$

4.2.1 Loss functions

As concluded in the previous chapter, the goal in model training is to minimize a given loss function on the training data (equation 4). Most famous of all loss functions is the linear least squares (LS) loss,

$$\Phi = (\mathbf{y} - \hat{\mathbf{y}})^2. \tag{5}$$

In LS algorithm, model predictions $\hat{\boldsymbol{y}}$ are simply subtracted from ground truth $\boldsymbol{y}$ and squared. Using linear models, estimate $\hat{\boldsymbol{y}}$ is calculated from product of features $\boldsymbol{X}$ and model coefficients $\boldsymbol{w}$, as $\hat{\boldsymbol{y}} = \boldsymbol{Xw}$. Averaging the LS loss (equation 5) over the included subjects, results in mean squared error (MSE). This is the empirical risk of LS loss and it makes the comparison between datasets of different sample size easier. Optimal solution is the one that finds the global minimum of the cost function on the parameter surface. Dimensionality of the parameter surface is equal to the feature dimension. [65]

Minimum solution can be obtained from setting the gradient of the cost function to zero. Considering the linear case, we can set the gradient of the MSE (mean of LS cost) to zero, computing the gradient along the parameter surface. After some equation solving, this results in optimum solution (equation 6)

$$\boldsymbol{\Sigma}_x \boldsymbol{w}_{opt} = \boldsymbol{p}, \tag{6}$$

where $\boldsymbol{\Sigma}_x$ is the covariance matrix of the features $\boldsymbol{X}$, $\boldsymbol{p}$ is the cross-correlation vector of features and ground truth, and $\boldsymbol{w}_{opt}$ is a vector containing the optimal model coefficients. [65]

However, finding the covariance matrix and cross-correlation vector is often not trivial and they have to be estimated. Optimization algorithms, such as steepest descent, stochastic gradient descent and conjugate gradient method can be used to iteratively find the solution of minimum error. Basis of gradient descent algorithms is simple to understand. Values of the loss function on parameter space form an error surface and optimum solution is found in the lowest possible point of the surface. The shape of this surface is not known unless evaluating the loss, and evaluating every possible point is computationally too expensive. Instead, computing the gradient of the loss shows the steepness of the error surface and quickest way to the bottom of the surface is through the steepest descent. Thus, next set of parameters should be evaluated from direction of the slope. In order not to divert from the path of steepest descent, the gradient has to be re-evaluated with a selected frequency to correct the direction. This frequency is known as the learning rate. [65]

4.3 Regression

Regression methods are used in supervised learning to predict continuous variables, such as severity of a disease. Often multivariate linear regression models are used. They may consist of multiple variables $\boldsymbol{X}$ multiplied by model coefficients $\boldsymbol{w}$. If one of the indices in $\boldsymbol{X}$ are set as 1, the term

equals to bias added to the model. Models should always contain an error residual ε. Thus, linear model can simply be formulated as in equation (7):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{w} + \epsilon \,. \tag{7}$$

Linear models are often optimized to minimal risk based on mean squared error (MSE). In MSE (equation 8), model prediction $\hat{\boldsymbol{y}}$ is assessed against ground truth $\boldsymbol{y}$:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \,. \tag{8}$$

Ridge regression (L2) was already introduced in 1970 by Hoerl & Kennard [68]. In the paper, they showed that adding small positive value (bias) to each term in least squares regression reduces errors in the prediction by reducing non-orthogonality of the data.

Logistic regression, despite its name is also a linear model. However, it can be used in binary classification. It can be also coupled with regularization to decrease errors [69]. For couple of years, linear regression methods have been used in pattern recognition applications [70].

4.4 Bias-Variance tradeoff and overfitting

In chapters 4.2 and 4.3, methods for creating an optimal model for a single dataset were discussed. Most of the time preferred models should be able to perform well on future data also. The error related to the model performance is composed of bias and variance. Bias is the difference between expected value of the model and mean of the ground truth. Variance of the model is the variance of the prediction mean. Variance of the test data is a third component that is always present on the real data and cannot be controlled.

Bias and variance of the model can be adjusted based on model complexity. Increasing the complexity usually increases the variance and decreases the bias of the model, resulting in a tradeoff between the two. A very complex model could find a combination of parameters that obtain a very low training error (e.g. MSE). However, the complex model has high variance, and is probably not going to give similar error on the test data, that the model has not seen previously. In this case, model overfits and does not generalize to unseen data due to too high adaptation to the training set. In the opposite case, a simple model has high bias, in which case both the training and test error are high due to underfitting. In the ideal case, bias and variance are balanced so that best performance can be achieved on the test data. [71]

4.4.1 Feature selection

Learning machines are trained to make predictions on features. These can include measurements from anything related to the assessed problem, e.g. blood pressure, texture patterns or gene sequences. To control the overfitting of trained models, the amount of features might require attention. If number of features used to train the model is more than the number of samples (high feature dimensionality), there is a high risk for overfitting. A model that is overfitted can easily find a combination of features that estimate the given data. Instead of learning only the variance of the data, model starts to predict noise in the training data. Consequently, the trained model performs well only on the training set and is useless when measuring completely new samples. [72]

To reduce chances of overfitting, dimensionality of the features can be reduced. Additionally to excluding irrelevant features, principal component analysis (PCA) can be used [73-75]. In PCA, feature array is re-projected in a coordinate system, where each component is linearly uncorrelated and the first component explains largest proportion of data variance possible. Other components are sorted in the order of highest explained variance. Thus, components with low explained variance can be omitted and lower number of components can be selected compared to including all of the features.

Example feature extractor: Local binary patterns

Local binary patterns (LBP) are textural features, first established by Ojala et al [76]. They have been shown to associate with OA features in subchondral bone [77]. Currently on computer vision applications, Median Robust Extended Local Binary Pattern (MRELBP) [78] is recognized as an efficient method for feature extraction in various applications [79].

Local Binary Patterns encode spatial grayscale environment of image pixels into a single value. In this simple but highly effective method, each image pixel is compared against its neighboring pixels (typically 8 neighbors N are used). If the neighbor x_n has higher gray-level intensity than center pixel x_c, a binary label 2^n is recorded. There are 2^N possible combinations when this comparison is made against all neighbors. Combination can be recorded as a single value using equation (9):

$$LBP(x_c) = \sum_{n=0}^{N-1} \text{sign}(x_n - x_c) \cdot 2^n, \quad \text{sign} = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} . \tag{9}$$

To reduce feature dimensionality and increase pattern robustness, these values can be mapped into a rotation invariant and uniform representation (riu2). This means combining all different orientations of the same pattern (rotation invariance) and detecting all patterns that do not include a "gap" in the pattern (uniformity). If no 0 → 1 transitions are allowed, only patterns 00000000_2 and 11111111_2 can be included. However, most often two transitions (leading to use of 2 in the abbreviation) are allowed, leading to N + 1 possible combinations. The non-uniform patterns are collected into one more bin and a total of N + 2 features are collected with riu2.

To summarize, instead of trying to incorporate as much features as possible to the model, it should be carefully weighed, which features are most descriptive of the studied phenomenon and should be selected for the model. To further reduce the risk of overfitting, dimensionality reduction is a possible choice. Low amount of descriptive features facilitates model generalization on inference.

4.4.2 Regularization

Regularization is a tool used in the model optimization to limit the solution from the absolute minimum value (and reduce overfitting). It was first introduced by a Russian mathematician Andrey Tikhonov for integral equations [80] and has been later implemented in statistics and linear models. Too high regularization starts to decrease the results and too small regularization does not compensate for the overfitting. The optimal regularization strength cannot be found analytically and has to be experimented upon each model [65].

With high parameter dimension, the LS algorithm has high risk of overfitting to a solution that does not generalize to unseen data. To reduce the chances of overfitting, different regularization protocols have been introduced [81]. Ridge- and Least Absolute Shrinkage and Selection Operator (LASSO) regression [82,83] models introduce a stabilizing bias term to the model. On ridge regression [68], bias term is the Euclidean L2 norm $\lambda||\boldsymbol{w}||^2$, where λ is a coefficient controlling the strength of the regularization. LASSO regression introduces a similar L1 one norm $\lambda||\boldsymbol{w}||$ without squaring the coefficients. Bias terms effectively decrease the relevance of small coefficients, i.e. coefficients that have only a small contribution to the result of the model.

4.4.3 Training and validation

As mentioned previously, choice of the samples in training the model might influence the final results. Optimizing large number of parameters or high feature dimension make it possible for the model to drastically reduce the error metric on the training data. This easily results in underestimation of the error metric (overfitting). On the other hand, simplifying the model too much leads to underfitting, e.g. using only the bias term in making the predictions.

Optimization of the parameters is always done on the training set. In the training process, besides reducing the error metric, the goal is to find combination of parameters that can provide a good estimate also to completely new samples. Utilizing the trained model on new data is called inference: the model infers and draws conclusions based on previous training on an unseen sample. Generalization of the model can best be seen on the inference phase. To estimate the actual performance of the model, some samples are always left out and are used as a test set that does not participate in the training phase. This is known as validation of the trained model.

There are multiple different validation strategies that can be utilized. When sample size is limited, cross-validation [84] is often used. In cross-validation, samples are split into k sets. Training is performed so that one set is left out for testing and $k - 1$ are used to train the model. This iterated through k number of folds, leaving out a different set for testing, until all sets are tested. After the iterations, the trained models as well as the error metrics obtained can be combined (averaging, median, etc.) to obtain the final estimates of the model and error metric. The method is known as k-fold validation, and it has advantage of utilizing all samples in training the model, while conducting testing that can be seen as independent.

A special case of k-fold cross-validation is using folds equal to number of samples ($k = n$). This method is known as leave-one-out (LOO) cross-validation [72]. This increases the number of iterations k, but larger portion of samples can be used in training the model on each fold. On some biomedical applications, to reduce overfitting due to overlap / autocorrelation of different samples, leave-one-patient-out (LOPO) –split can be used, excluding one patient on each fold. The possible overlap between training and testing folds is the largest downside of the cross-validation methods and if possible, an independent test set with large enough sample size should be included [85].

4.5 Assessment of predictive performance

MSE is one of the most popular metrics in estimating performance of different predictive models [65]. However, this single value does not necessarily tell everything from the model, and more extensive methods can be used to further assess the performance. For linear regression models, Spearman's correlation along with the p-value is a good method to assess the linear dependency of the predictions and ground truth. Plotting predictions against the ground truth allows visual assessment how realistic the predictions are, and outliers can be easily detected. Other methods include Wilcoxon's signed rank test or coefficient of determination (R^2 score).

Providing confidence intervals for error metrics gives information from the metric variance. One method for this is bootstrapping. It is a data resampling method based on sample replacement, and it allows increasing the apparent size of the used dataset. Bootstrapping drastically increases the reliability of the analysis, especially for relatively small datasets. On biased data, stratified bootstrapping should be used, sampling equally from both classes. [72]

Most of the performance evaluation methods of classifier models are based on analyzing variables of the confusion matrix (Figure 4). True positive rate (sensitivity, recall), false positive rate and accuracy are often reported together. Other informative values are precision (positive predictive value), specificity (true negative rate) and F1 score (combined from precision and recall). Other common combinations that are reported are precision-recall and sensitivity-specificity.

Confusion matrix

	Positive **(P)**	Negative **(N)**
Predicted positive	**True positive (TP)**	**False positive (FP)**
Predicted negative	**False negative (FN)**	**True negative (TN)**

Statistical variables

False positive rate	$\boldsymbol{\frac{FP}{P}}$	True positive rate, recall, sensitivity	$\boldsymbol{\frac{TP}{P}}$
True negative rate, 1 - specificity	$\boldsymbol{\frac{TN}{N}}$	False negative rate	$\boldsymbol{\frac{FN}{N}}$
Precision, positive predictive value	$\boldsymbol{\frac{TP}{TP+FP}}$	Accuracy	$\boldsymbol{\frac{TP+TN}{P+N}}$
F1 score	$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$		

Figure 4. Illustration of the confusion matrix and variables derived from it. Most commonly used names for different statistical variables are listed. High performance classifiers have most samples situated on the confusion matrix diagonal.

When the model is able to provide a probability estimate or a range of values related to the prediction (for example logistic regression and random forest), the binary statistical variables have to be acquired using a threshold for the binary classification, e.g. 50%. These models can be evaluated more extensively, using the full range of possible thresholds. Informative plots can be produced using combinations of the binary metrics. Receiver operating characteristic (ROC) curve is a true positive rate-false positive rate –graph and area under the ROC curve (AUC) is reported often from the analysis [86]. Sensitivity-specificity –curve is very similar to ROC curve, since specificity is 1 – true negative rate. Precision-recall curve (PRC) is a third curve produced in similar manner (using different variables), with average precision (AP) as the metric that can be obtained from the curve. PRC curves have been shown to better evaluate biased datasets, where distribution of the classes is not even [87].

To summarize, there is a very large toolbox of different methods for evaluating performance of the predictive models. There is a separate set of tools for regression and classification problems that can be used. To select which metrics to use, it is good to provide an informative metric of the prediction error along with a measure of statistical significance. Special care should be taken to avoid overestimation of the model performance in case of biased data and internal validation. Independent testing should always be conducted with sufficient amount of samples when possible.

5. Materials and methods

5.1 Samples

Human osteochondral cores were harvested from 24 TKA patients and 2 asymptomatic cadavers. Two sample sets were selected in this study based on core diameter (Cross-validation set; n = 34, $N_{patients}$ = 19, Ø = 2 mm, ethics approval PPSHP 78/2013, The Northern Ostrobothnia Hospital District) (Test set; n = 45, $N_{patients}$ = 7, Ø = 4 mm, ethics approval PPSHP 78/2013; PSSHP 58/2013 & 134/2015, The Norther Savo Hospital District). Distribution of patients and core extraction locations are given in Table 1.

During surgery, tissue blocks from the femoral head and tibial plateau were extracted and stored in 1% phosphate buffered saline (PBS) solution. Cores were extracted from the blocks using a dental drill (biopsy head sizes: Ø = 2 mm, Ø = 4 mm). Cartilage drying was prevented by spraying PBS during core extraction. Samples were stored in PBS solution and frozen in -80°C until the time for CEμCT imaging. After thawing, samples were fixed for 5 days in 10% neutral buffered formalin to preserve sample microstructure and composition. Subsequently, samples were immersed in 70% ethanol 1% w/v PTA solution (1g of PTA / 100ml of solution) to label the collagen content of AC for μCT imaging. For the imaging, each sample was wrapped in Parafilm (Parafilm M, Bemis Company Inc, Neenah, WI, USA) and orthodontic wax (Orthodontic Wax, Ortomat Herpola, Turku, Finland) to prevent drying during μCT scan.

5.2 CEμCT imaging

After 48h PTA immersion samples were imaged with a desktop μCT system (Skyscan 1272; Bruker microCT, Kontich, Belgium; 45 kV, 222 μA, 3.2 μm voxel side length, 3050 ms, 2 frames/projection, 1200 projections, 0.25 mm aluminum filter) and reconstructions were calculated from projection images using the manufacturer's software (NRecon version 1.6.10.4; Bruker microCT, Kontich, Belgium) with beam-hardening and ring artefact corrections applied. Reconstruction algorithms were based on cone-beam FBP by Feldkamp et al [40].

Some of the samples in the Cross-validation set included a "void" [51]. This refers to a volume inside the deep cartilage layer of the scans, with no PTA accumulation present due to insuf-

ficient diffusion. In the Test set, possible voids were avoided by increasing the PTA immersion time.

Table 1. Distribution of µCT grades assessed from reconstructions (used as ground truth). Cross-validation set contained only small number of samples from grade 3 and reduced number of healthy samples, while almost no healthy samples were found in the Test set. Otherwise samples were distributed quite evenly. Core extraction location, number of samples and patients as well as TKA / Cadaver status are described.

Dataset	Zone	Grade 0	Grade 1	Grade 2	Grade 3
Cross-validation set					
Only TKA patients Total: n = 34, $N_{patients}$ = 19 Tibia: n = 16, $N_{patients}$ = 16 Femur: n = 18, $N_{patients}$ = 18	Surface	**7**	**11**	**13**	**3**
	Deep	**8**	**16**	**8**	**2**
	Calcified	**8**	**16**	**7**	**3**
Test set					
Only tibial cores Total: n = 45, $N_{patients}$ = 7 TKA: n = 15, $N_{patients}$ = 5 Cadaver: n = 30, $N_{patients}$ = 2	Surface	**2**	**19**	**9**	**14**
	Deep	**0**	**16**	**15**	**13**
	Calcified	**0**	**24**	**11**	**9**

5.3 3D Histopathological grading

After CEμCT imaging, ground truth was evaluated for different osteochondral zones using previously developed 3D grading system [6]. Different zones of the osteochondral samples were manually graded based on visual inspections of the full volumetric reconstructions. Grade distributions are illustrated in Table 1 and visual examples of the reconstructed slices in Figure 5. In this study, we used the following 3D CEμCT grades as the ground truth:

- Surface discontinuity (SZ, Smooth and continuous = 0; Slightly discontinuous = 1; Moderately discontinuous = 2; Severely discontinuous = 3)
- Deep cartilage ECM disorganization (DZ, Normal = 0; Slightly disorganized = 1; Moderately disorganized = 2; Severely disorganized = 3)
- Calcified cartilage ECM disorganization (CZ)

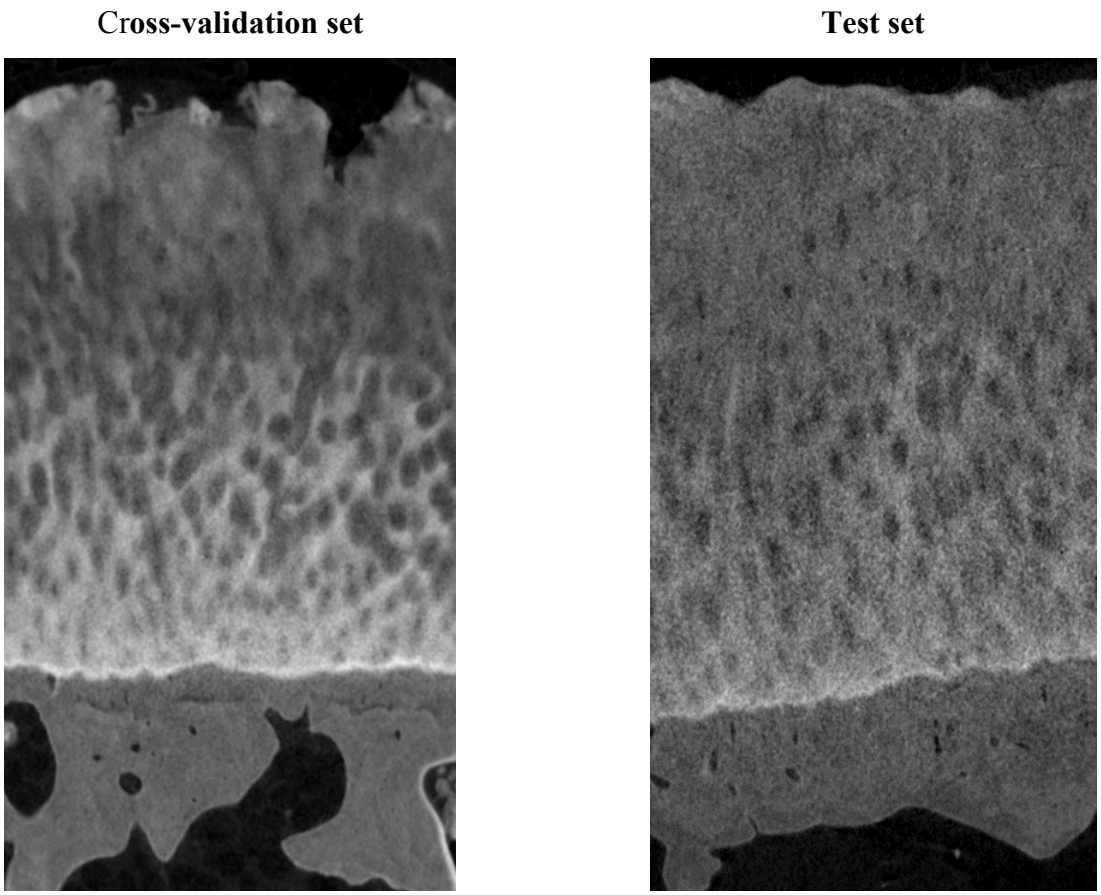

Figure 5. Example slices from samples in Cross-validation and Test sets. Image quality is visually higher in the Cross-validation set. However, osteochondral features such as chondrons and vascular channels can be distinguished visually in both samples.

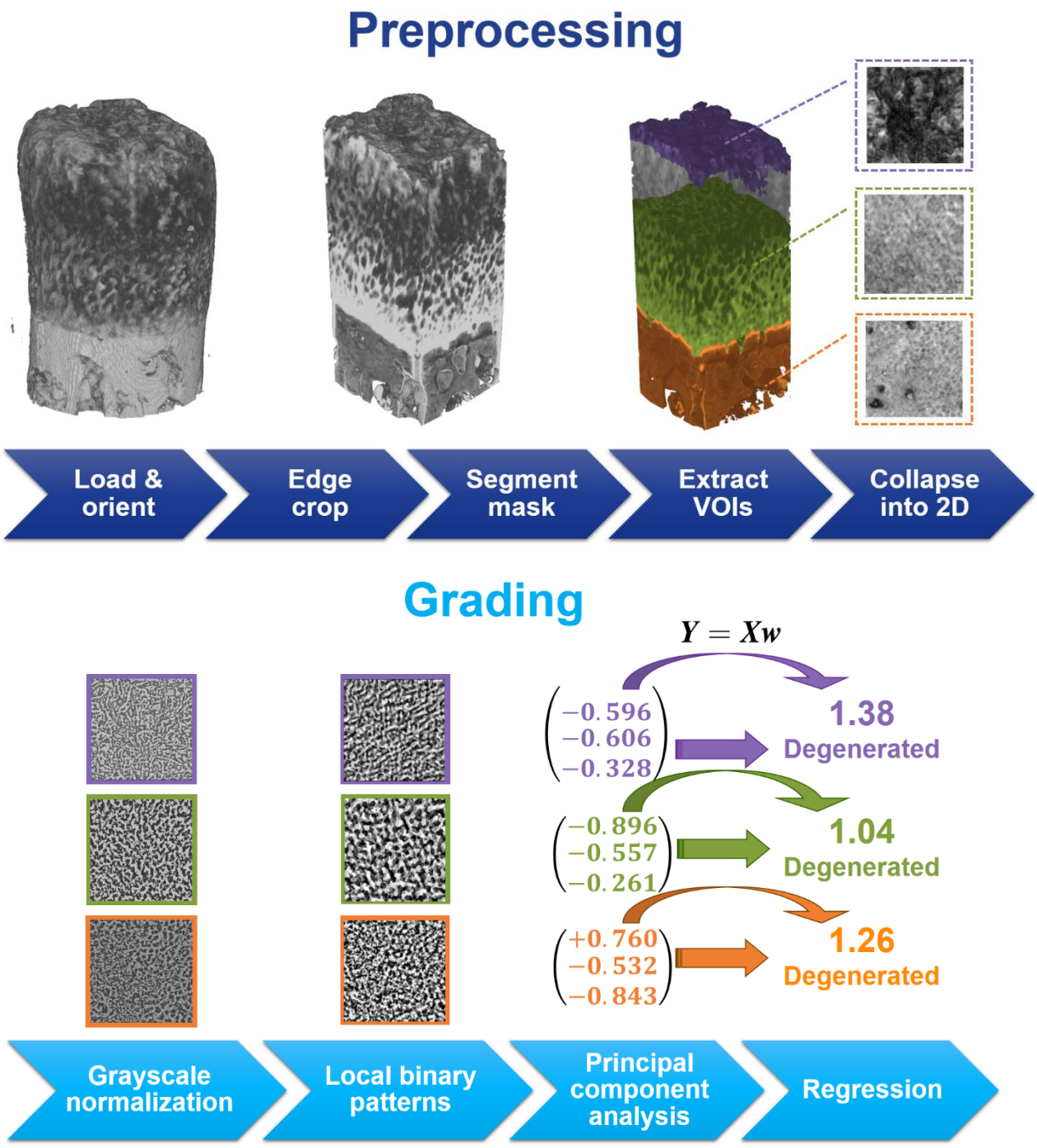

Figure 6. Workflow of the analysis methods used in the Python and prototype software. In the preprocessing pipeline, CEμCT imaged osteochondral samples are loaded, oriented and edge cropped. Calcified tissue is segmented, VOIs are extracted and collapsed into 2D. In the grading pipeline, images are normalized, MRELBP and PCA are calculated, and finally Ridge and Logistic regression models are used to obtain predictions of each VOI's degeneration. Adapted from [1].

5.4 CEμCT data preprocessing

To automate the grading process, a python software was developed. Python software functionalities run on computer central processing unit (CPU) and most time-consuming calculations are parallelized to allow shorter processing times. Orthogonal planes and volume renderings can be saved from each processing step to verify the performance of the software and detect samples that are not suitable for analysis (e.g. decalcified samples, or samples with either no bone or cartilage present). Workflow of the grading software is shown in Figure 6.

In the preprocessing and VOI extraction pipeline (Figure 6, top), reconstructed samples are loaded from given data path and sample borders from each slice are estimated using a bounding rectangle fit. Samples are then oriented using one of the developed orientation algorithms. Bounding rectangle algorithm calculates linear fits based on rectangle coordinates to estimate sample orientation. Principal component analysis (PCA) can be used to find the sample's main axis. This method works only on sufficiently tall samples, where sample axis is evident. In this study, we used a circle-fitting algorithm that compares depth-wise sum of the sample to a circle fitted in the sum projection, in a gradient descent –based optimization loop.

After orientation, a depth-wise sum image was calculated, and sample center was estimated using grayscale-weighted average coordinates of the sample. Center coordinates were used to crop off the edges of the samples resulting in rectangular volumes (1300μm · 1300μm · Z for the Cross-validation set, 2600μm · 2600μm · Z for the Test set).

Non-calcified AC was segmented from calcified tissues after the edge crop. For the Cross-validation set, a convolutional neural network (CNN) was used trained with 5-fold cross-validation, using μCT scans without PTA contrast as the ground truth. Ground truth masks were co-registered to PTA scans using rigid transformations. Segmentation was performed slice-by-slice using a U-Net [88]. Segmentations on XZ and YZ plane were averaged and mask was created thresholding by 0.5. The segmentation model used is a pretrained network from an upcoming paper by Tiulpin et al. (unpublished during the writing of this thesis). Reason to use CNN segmentation was to take the voids in the Cross-validation set into account, and accurate segmentations could be provided. Regarding the voids, the imaging protocol was improved for the Test set by checking the voids immediately after μCT scanning. If a void was found, sample was immersed again in PTA to allow diffusion to full cartilage volume, followed by rescanning the sample with μCT.

The trained CNN model did not generalize to the Test set with sufficient accuracy, and another, test set-specific algorithm was developed based on k-means clustering to segment the

tidemark. Due to the absence of voids in the Test set, cluster with highest average intensity was used as a boundary for deep cartilage. Segmentation was performed also slice-by-slice and area above the bottom of deep cartilage cluster was labeled as non-calcified cartilage.

After calculating the calcified tissue mask, volumes-of-interest (VOI) corresponding to SZ, DZ and CZ were extracted from the sample. Depth of AC was estimated using the mask and the thresholded cartilage surface, averaged from the cropped VOI along the XY-plane. Depth of SZ was set to 160µm (excluding background in case of complex structure), 60% of calculated cartilage depth for DZ and 160µm for CZ. Surface was located based on Otsu thresholding. DZ and CZ were adjacent with the border of zones being set 30µm above the segmented interface. To encode the volumetric 3D information into 2D image, mean and standard deviation of the gray-level values were calculated along cartilage depth for each VOI. Results were summed, encoding the volumetric information of the reconstruction into a 2D texture image.

For the larger samples in the Test set, nine smaller subimages were created (1300µm · 1300µm) with equal spacing on the large texture image. This resulted in slightly overlapping images with similar size compared to the Cross-validation set. Smaller images were used to increase prediction reliability and scale the relative size of detected MRELBP features similar to the Cross-validation set.

5.5 Grading of extracted texture images

In the grading pipeline for the texture images (Figure 6, bottom), possible artefacts on DZ and CZ image were first cropped out using an automatic algorithm. Adaptive thresholding was used to detect artefacts on image edges due to misalignment or issues with segmentation. Subsequently, the images were normalized to local contrast using Gaussian normalization. Kernel sizes and standard deviations σ (σ^2 refers to the variance) for equation (10) were optimized (Table 2) and convolution was applied using defined kernels to estimate local mean and standard deviation. 2D kernels *G* were created implementing the normally distributed Gaussian function

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \quad . \tag{10}$$

Here, x and y refer to distance from center of the kernel. Local mean image was subtracted from the input image and local standard deviation image was divided from the subtraction result.

5.5.1 Local Binary Patterns

After normalization, images were used to calculate MRELBP images (Figure 7) according to Liu et al [78]. Parameters that were used, are listed on Table 2. Four images were created: Center LBP (LBP_C), Large LBP (LBP_L), Small LBP (LBP_S) and Radial LBP (LBP_R). Center, Large and Small image were initialized by median filtering the normalized input image according to kernel sizes listed. Average value μ_c from center image was calculated and subsequently subtracted element-wise. From this image, two features were calculated for the final MRELBP histogram. These were the sum of image pixels $LBP_C(x_c)$, whose value was < 0 and pixels with values ≥ 0. Value of each pixel is obtained as described in equation (11)

$$LBP_C(x_c) = \text{sign}(\mathcal{F}(\boldsymbol{Im}(x_c)) - \mu_c) \ , \tag{11}$$

where $\mathcal{F}$= corresponding median filter and ***Im*** = normalized input image array.

To obtain LBP_L and LBP_S, each image pixel x_c was compared against its neighbors x_n on circular radius specific for both LBP images. This was conducted by calculating neighbor-images $\boldsymbol{Im}_n$ with corresponding distance from the center of the filtered image. This way, pixel in coordinates $\boldsymbol{Im}_n(x_c)$ corresponds to distance of the radius from pixel in $\boldsymbol{Im}_n(x_c)$. Value of diagonal neighbors was estimated using bilinear interpolation. To allow comparison between neighbors, resulting image size is $2 \cdot$ large radius smaller on both dimensions. LBP images were obtained by subtracting mean of the median filtered neighbor-image μ_n element-wise and attaching values ≥ 0 with the binary label 2^n. This is formulated in equation (12), corresponding to each pixel x_c as

$$LBP_{L,S}(x_c) = \sum_{n=0}^{N-1} \text{sign}(\mathcal{F}(\boldsymbol{Im}_n(x_c)) - \mu_n) \cdot 2^n . \tag{12}$$

Finally, LBP_R was created simply by subtracting gray-values of median filtered neighbor pixels x_n of both radiuses and attaching the binary label 2^n to values ≥ 0 as seen from equation (13)

$$LBP_R(x_c) = \sum_{n=0}^{N-1} \text{sign}(\mathcal{F}(\boldsymbol{Im}_{n,L}(x_c)) - \mathcal{F}(\boldsymbol{Im}_{n,S}(x_c))) \cdot 2^n \ . \tag{13}$$

Subsequently, riu2-mapping was used for LBP_L, LBP_S and LBP_R to reduce dimensionality from 256 to 10. All four histograms were concatenated and total of 32 features were obtained. This feature vector was normalized by dividing each element with sum of the vector. Features that had zero occurrences were removed, resulting in 28 remaining features.

Features were obtained for all samples in the Cross-validation set and a mean feature was calculated. Each feature vector was centered by subtracting the mean feature. Along with

created model, the mean feature was saved to allow centering the Test set features (and other features, when conducting an inference).

Table 2. Parameters optimized in contrast normalization and MRELBP, attached with description. Normalization parameters are shown on top of the table, while LBP parameters are on the bottom. For SZ, DZ and CZ models, the same parameter set was used. However, a second set occurred frequently during optimization of CZ model. Adapted from [1].

Parameter	Values used	Frequently encountered values in CZ (16/34)	Description
Centering kernel size	**25**	**23**	Gaussian kernel size for centering the extracted texture image (subtracted from input)
Standardizing kernel size	**21**	**21**	Gaussian kernel size for standardizing the input image (divided from centered image)
σ_{mean}	**4**	**4**	Standard deviation of centering Gaussian kernel
σ_{std}	**7**	**6**	Standard deviation of standardizing Gaussian kernel
Neighbors	**8**	**8**	Number of neighbors used in MRELBP (4 orthogonal and 4 diagonal neighbors).
LBP_L radius	**18**	**12**	Circular distance of center pixel from neighbors used in obtaining large image
LBP_S radius	**4**	**11**	Circular distance of center pixel from neighbors used in obtaining small image
Center kernel LBP_C kernel (median filter)	**15**	**9**	Kernel size used for median filtering center image
LBP_L kernel (median filter)	**15**	**9**	Kernel size used for median filtering large LBP image
LBP_S kernel (median filter)	**13**	**15**	Kernel size used for median filtering small LBP image

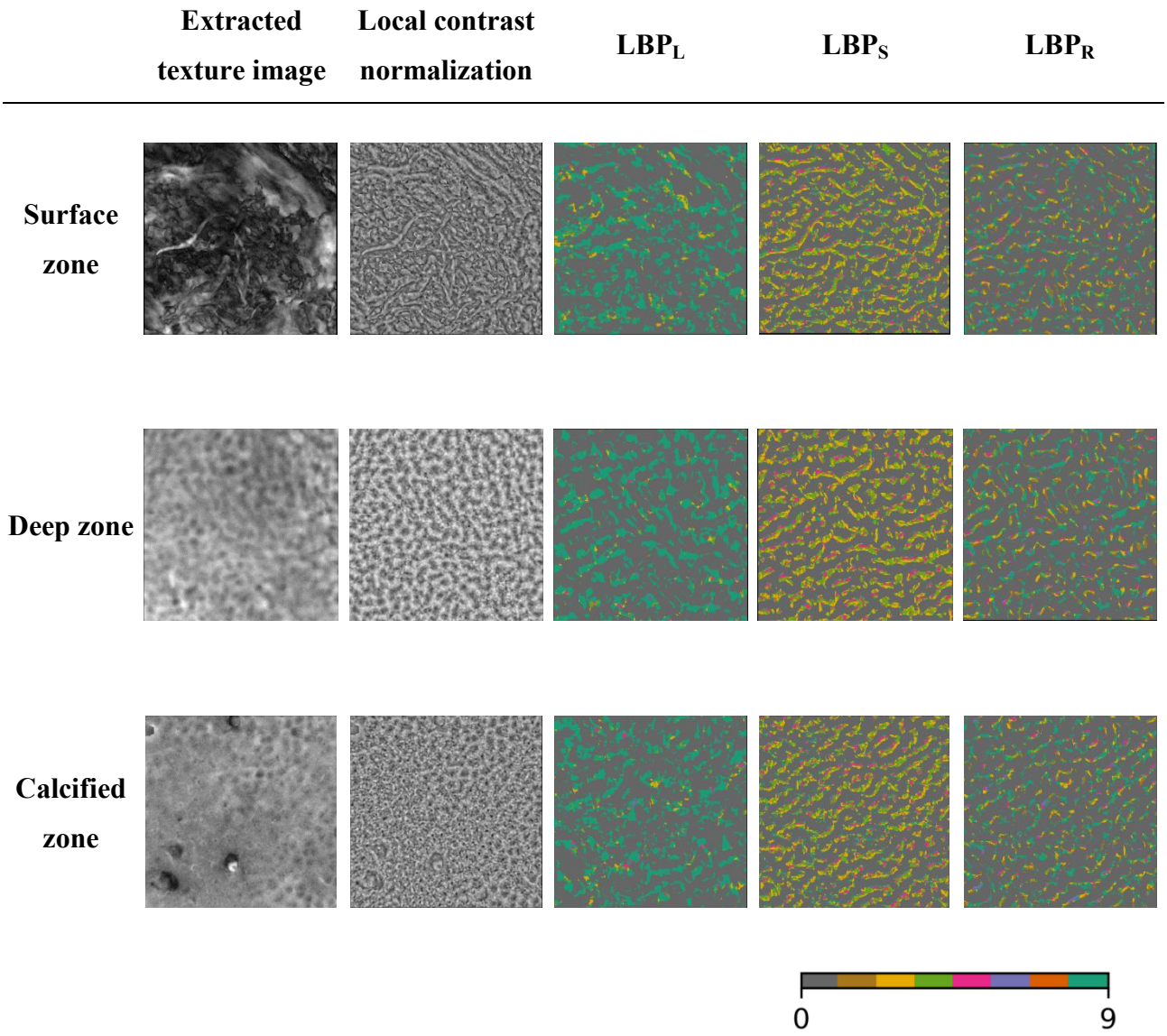

Figure 7. Feature extraction steps applied for the 2D texture images. Calculated mean + standard deviation texture image, normalized image, as well as LBP_L, LBP_S and LBP_R are shown. Colorbar is used to distinguish the different features obtained through riu2-mapping.

5.5.2 Principal component analysis

To conduct PCA dimensionality reduction, full singular value decomposition (SVD) was used with implementation from Linear Algebra Package [89]. This is done by reducing first the feature array $\boldsymbol{A}$ to bidiagonal form $\boldsymbol{A} = \boldsymbol{U_1 B V_1^T}$. Arrays $\boldsymbol{U_1}$ and $\boldsymbol{V_1}$ are orthogonal and $\boldsymbol{B}$ is the bidiagonal array (array where indices are nonzero only on main diagonal and either on diagonal above or below main diagonal). Afterwards, SVD is computed for the bidiagonal $\boldsymbol{B}$, resulting in $\boldsymbol{B} = \boldsymbol{U_2 \Sigma V_2^T}$. $\boldsymbol{U_2}$ and $\boldsymbol{V_2}$ are orthogonal arrays and $\boldsymbol{\Sigma}$ includes the singular values in descending order. Left and right singular vectors are obtained by multiplication $\boldsymbol{U} = \boldsymbol{U_1 U_2}$ and $\boldsymbol{V} = \boldsymbol{V_1 V_2}$, respectively.

PCA components are obtained as columns of matrix $\boldsymbol{V}$ using 90% of total explained variance of the components (from feature array $\boldsymbol{A}$). Explained variance of the singular vectors can be calculated as $\boldsymbol{\Sigma}^{\boldsymbol{2}}$ / *(number of samples + 1)*. Whitening of the PCA components can be performed by multiplying with $\sqrt{n_{samples}+1}/\boldsymbol{\Sigma}$ the term $W = \sqrt{n_{samples}+1}/\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is also truncated to 90% explained variance. Whitening sets the variance of the components to 1.

In the Python software, PCA components were calculated for the centered features of the Cross-validation set using SVD algorithm described above and whitening was applied. PCA components $\boldsymbol{V}$ and the whitening term W were saved in the resulting model for evaluation of the Test set.

To perform inference on the Test set, PCA was applied using matrix multiplication between the feature array $\boldsymbol{A}$ and PCA components $\boldsymbol{V}$. Result was multiplied with W to apply the whitening. Components from the zones of the Cross-validation and Test set are plotted in Figure 8, where the samples with low and high degeneration are separated.

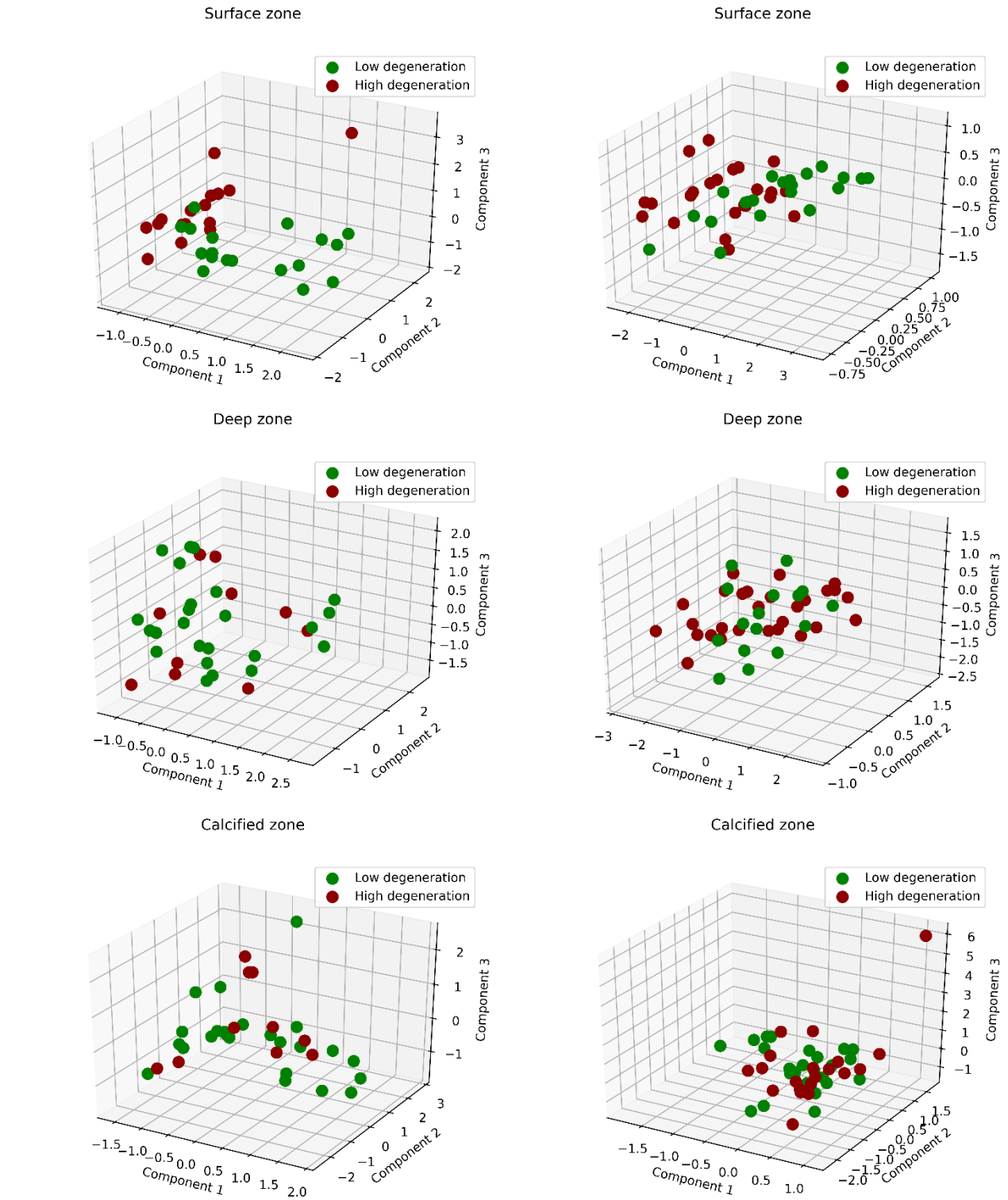

Figure 8. Extracted PCA components from the Cross-validation and Test sets. Samples with grades ≤ 1 are labeled green (Low degeneration) and grades > 1 are labeled red (High degeneration). Groups are visually best separated in SZ (both datasets), and DZ of the Test set.

5.5.3 Regression

In the software, we created two models: a ridge regression model [68] and logistic regression model. Ridge regression model was trained for the PCA components ***V*** against ground truth ***y*** optimized with linear least squares, regularized with L2 norm (Tikhonov regularization), using loss function Φ according to equation (14)

$$\Phi = \|\mathbf{y} - \mathbf{V}\mathbf{w}\|_2^2 + \alpha \|\mathbf{w}\|_2^2 , \tag{14}$$

where ***w*** is a vector containing model weights (and bias for dummy feature = 1) and α is the regularization coefficient. $\|\cdot\|_2^2$ refers to the Euclidean L2 norm. Logistic regression model was optimized with Newton conjugate gradient (Newton-CG) –method and L2 regularization. The loss function can be formulated as

$$\Phi = C\sum_{i=1}^{n} \log(e^{-y_i(V_i^T w + c)} + 1) + \min_{w,c} \frac{1}{2}\mathbf{w}^T\mathbf{w} . \tag{15}$$

In equation (15), *i* is an index iterated through the training samples *n*, *C* controls the regularization strength and *c* is a residual term. In the last term, minimum of either the residual term or the L2 norm is added as a bias. In this notation, ground truth observations are labeled as -1 and 1.

Both models were trained on the Cross-validation set using leave-one-patient-out (LOPO) cross-validation. Coefficients and intercept of both models were saved for evaluation on the Test set. Evaluation on the Test set was performed by multiplying the estimated PCA components of each subimage with the coefficients and adding the intercept. Final prediction on the Test set was estimated using average of the results on the nine subimages.

5.6 Parameter optimization

Hyperparameter search is a common problem in numerical method optimization. Often used methods for hyperparameter search are random and grid search. In grid search, each variable is given values on an equally spaced grid and every possible combination is evaluated. When number of hyperparameters increase, computational demand becomes intensive, since a much larger parameter space has to be evaluated. However, evaluating random combinations often performs equally with decreased workload [90]. Compared to random and grid search, more sophisticated methods have been developed based on Bayesian optimization. One such method is the Tree of Parzen Estimators (TPE) [91,92] algorithm. TPE utilizes sequential Gaussian Mixture Models to find its optimal solution.

To obtain optimal normalization and MRELBP parameters in the Python pipeline, Tree of Parzen Estimators (TPE) –algorithm was used on the Cross-validation set. Optimization was performed using a "nested cross-validation" approach. Hyperparameter search was conducted on N – 1 samples for 34 iterations with LOO split. During every iteration, another cross-validation was performed using LOPO-split (nested cross-validation) to train the model and evaluate MSE loss of the predictions. Maximum of 100 parameter sets were evaluated / iteration. Resulting parameter sets were saved and the most frequent solution was taken from the 34 sets. Algorithm converged on same parameter set on all zones (30/34 for SZ, 34/34 for DZ and 18/34 for CZ), however CZ model had 16 occurrences of another parameter set (Table 2).

5.7 Statistics and performance evaluation

Linear regression models have long been evaluated based on MSE [68]. Spearman's correlation can be used to evaluate prediction dependency in ordinal variables such as the μCT grade. It is more resistant to outliers compared to Pearson correlation, and should always be used when dealing with relationship of ordinal variables. Realistic predictions should have somewhat linear dependency with the ground truth.

Binary classification has been often reported using receiver operating characteristic (ROC) curves [86]. In ROC curves, true positive rate (or recall or sensitivity) is plotted against false positive rate to assess model's sensitivity at the cost of classifying false positives. ROC curve analysis provides easy-to-understand graph that should be used solely when dataset is evenly distributed. However, if distribution is uneven (often there is an abundance of positives), this can

lead to wrong conclusions on the data. Precision-recall curves (PRC) are more descriptive on imbalanced data and should always be reported in such cases [87].

For binary classification, there are multiple metrics that can be evaluated. Area under the ROC curve (AUC) is often reported from the performance analysis. AUC of 1 resembles perfect classification and 0.5 is equal to random guess. Model can give AUC values lower than 0.5, but 1 - AUC can in some cases be used, when inverting the class labels. Precision is the proportion of true positives in all predicted positives. Recall (or sensitivity or true positive rate) is the proportion of true positives from all positive samples. Often the harmonic mean of precision and recall (F1 score) is also reported. Plotting precision against recall on every possible threshold outputs the PRC curve. From PRC analysis a very informative metric is the average precision (AP). It condenses performance of the binary model well into a single variable.

To estimate confidence intervals from values obtained from the ROC and PRC curves (AUC and AP), often used method is bootstrapping. This means evaluating the bootstrapped metric multiple times, while randomly sampling from the data. On imbalanced datasets, stratified bootstrapping should be used. Stratification means sampling from both classes (true or false) independently alleviating the limitations of imbalanced data [93].

Extensive statistical analysis was conducted to assess the predictive power of the models. Ridge regression models were evaluated using MSE and Spearman's correlation (p-value was calculated to assess statistical significance). Predictions were also assessed visually by evaluating scatter plots against the ground truth (Figure 9).

The overall performance of logistic regression models were evaluated using ROC curve and PRC analysis (Figure 10). From these plots, AUC and AP were calculated. Both metrics were evaluated for 95% confidence intervals using stratified bootstrapping (2000 iterations were used). Predictive power on a threshold of 0.5 was also analyzed using precision, recall and F1 score.

5.8 Replication experiment and data acquisition differences

To further test the feasibility of the presented texture-approach, a separate model was trained on the Test set. Internal cross-validation was used with LOPO-split. Separate model was compared to the inference of Cross-validation set model on Test set to see if internal training would improve the metrics.

Visual inspection of the reconstructed μCT images revealed that Test set contains visually much more noise compared to the Cross-validation set. To quantify this, three metrics were calculated against median filtered μCT images (kernel size 5): MSE (equation 8), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). PSNR can be calculated as

$$PSNR = 10 \cdot \log_{10}\left(\frac{\max^2}{MSE}\right), \tag{16}$$

where max = maximum pixel intensity value of the μCT image. SSIM has a slightly more complicated formula, and can be calculated as

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{17}$$

Where x denotes the reconstructed image, y the median filtered image, μ = average of the image and σ = variance of the image. σ_{xy} is the covariance of the two images. Division in the equation is stabilized using variables c_1 and c_2. Using equations (8), (16) and (17) MSE, PSNR and SSIM were calculated for coronal μCT images taken from each sample from multiple slices (~50/sample) with equal spacing and averaged to obtain one value / sample / metric.

6. Results

Table 3. Performance of the trained Ridge and Logistic regression models. Confidence intervals for 95% are given in the parentheses. Statistical variables for Ridge regression are on left side of the table and the variables for Logistic regression are on the right side. Strong and significant correlations, AUC, AP > 0.8, and precision, recall, F1 > 0.6 are bolded. Adapted from [1].

Dataset	Zone	Ridge Regression			Logistic Regression				
		MSE	SC	p-value	AUC	AP	Prec.	Recall	F1
Cross-validation	S	0.49	**0.68**	**< 0.0001**	**0.92** (0.80-0.99)	**0.89** (0.77-0.99)	**0.83**	**0.94**	**0.88**
	D	0.66	0.38	**0.02**	0.72 (0.54-0.88)	0.50 (0.35, 0.75)	0.44	**0.80**	0.57
	C	0.50	0.54	**0.001**	0.77 (0.54, 0.94)	0.71 (0.48-0.91)	0.41	**0.70**	0.52
Test	S	0.85	0.55	**0.0001**	**0.86** (0.73-.95)	**0.89** (0.78-0.96)	**0.78**	**0.61**	**0.68**
	D	1.30	0.34	**0.02**	0.72 (0.56-0.86)	**0.83** (0.73, 0.93)	**0.84**	0.57	**0.68**
	C	1.01	0.29	**0.05**	0.63 (0.45-0.78)	0.62 (0.48-0.77)	**0.62**	0.40	0.49

S = Surface zone, D = Deep zone, C = Calcified zone, SC = Spearman's correlation, Prec. = Precision, F1 = F1 score

6.1 Ridge regression models

Ridge regression model trained on optimized MRELBP and normalization parameters gave following results on calculated metrics on Cross-validation set (Figure 9, Table 3): MSEs were 0.49, 0.66 and 0.50 for SZ, DZ and CZ models, respectively. On the model for SZ, strong correlation was observed ($\rho = 0.68$). For the CZ model, moderate correlation was observed ($\rho = 0.54$) and for the DZ model, weak correlation was observed ($\rho = 0.38$).

Evaluation of the Ridge regression models on the Test set resulted in following: MSEs were 0.85, 1.30 and 1.01 for SZ, DZ and CZ models, respectively. Spearman's correlations were moderate ($\rho = 0.55$) on SZ model, and weak ($\rho = 0.34, 0.29$) on DZ and CZ models. Spearman's correlations on all datasets and zones were statistically significant.

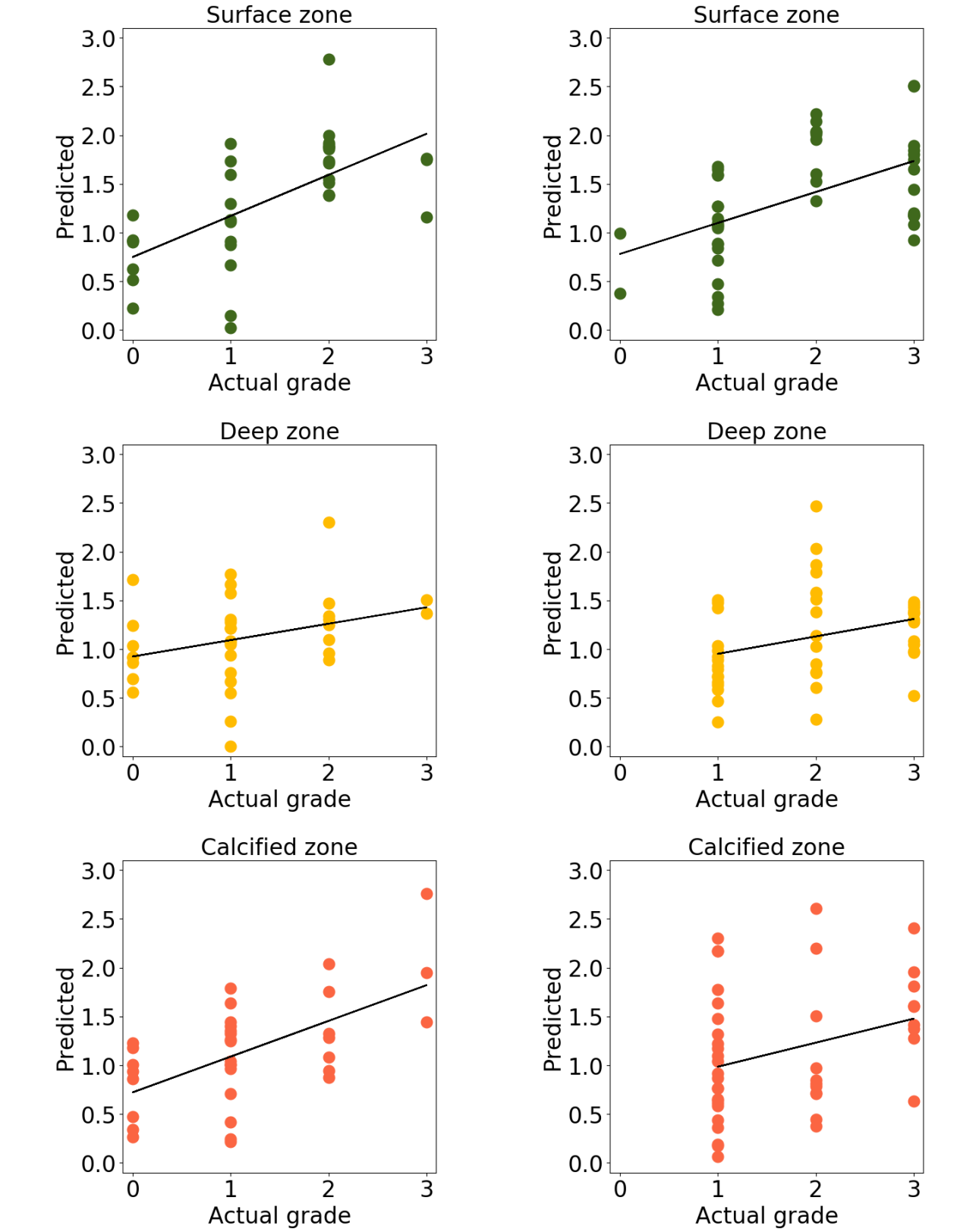

Figure 9. Predictions of Ridge regression models on Cross-validation and Test sets against the manually assessed ground truth. Adapted from [1].

6.2 Logistic regression models

For the Logistic regression model on Cross-validation set, we obtained following results (Figure 10, Table 3):

- **ROC curve analysis:** AUC values of 0.92 [0.80, 0.99], 0.72 [0.54, 0.88] and 0.77 [0.54, 0.94] were obtained for SZ, DZ and CZ, respectively.
- **PRC analysis:** APs of 0.89 [0.77, 0.99], 0.50 [0.35, 0.75] and 0.71 [0.48, 0.91] were obtained for SZ, DZ and CZ, respectively.
- **Confusion matrix analysis, threshold of 0.5:** precision values were 0.83 (SZ), 0.44 (DZ) and 0.41 (CZ). Recall (sensitivity) values were 0.94 (SZ), 0.80 (DZ) and 0.70 (CZ). F1 scores were 0.88 (SZ), 0.57 (DZ) and 0.52 (CZ).

Results of Logistic regression model on Test set were:

- AUC values of 0.86 [0.73, 0.95], 0.72 [0.56, 0.86] and 0.63 [0.45, 0.78] were obtained for SZ, DZ and CZ, respectively.
- APs of 0.89 [0.78, 0.96], 0.83 [0.73, 0.93] and 0.62 [0.48, 0.77] were obtained for SZ, DZ and CZ, respectively.
- For threshold of 0.5, precisions were 0.78 (SZ), 0.84 (DZ) and 0.62 (CZ). Recall values were 0.61 (SZ), 0.57 (DZ) and 0.40 (CZ). F1 scores were 0.68 for SZ and DZ and 0.49 for CZ.

Model for SZ showed comparable performance metrics on Test set compared to Cross-validation set. Performance decreased on CZ, while an increase was seen on DZ. On DZ model, AP value increased 0.33 from Cross-validation set.

ROC and PRC analysis (Figure 10) show that SZ is performing best of all zones. ROC curves on Cross-validation set show slightly better performance on CZ compared to DZ, while difference is much clearer on the PRC plot. Similar observations can be seen on Test set, except that the DZ model outperforms the CZ model.

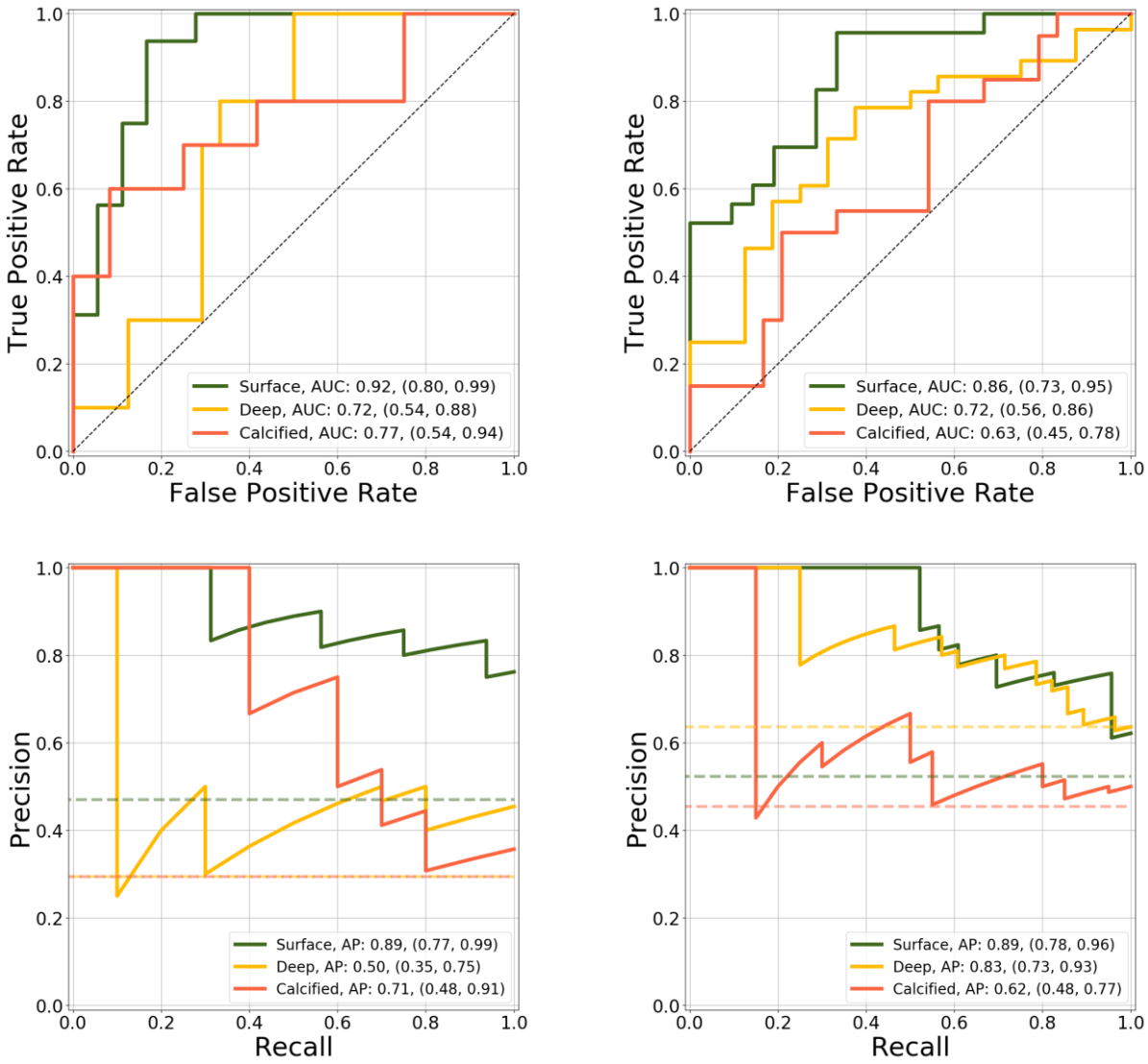

Figure 10. Performance of the Logistic regression models on Cross-validation and Test sets evaluated using ROC curve and PRC curve analysis. Adapted from [1].

6.3 Test set analysis

Separate model for the Test set was trained using internal cross-validation to assess the feasibility of presented texture analysis approach on the Test set. Results of the training are listed on Table 4. On Ridge regression model, separate training improved MSE of the predictions, but not the Spearman correlations. Logistic regression models for SZ and CZ reached comparable results to inference, apart for improvement of recall values for SZ and CZ on the threshold of 0.5. Logistic regression model for DZ was not improved due to training.

Table 4. Results for model trained on Test set with internal cross-validation. MSE values on ridge regression are improved, but Spearman correlation values do not show improvement. Logistic regression shows comparable performance on SZ and CZ, but lower on DZ. Adapted from [1].

Zone	Ridge regression			Logistic Regression				
	MSE	**SC**	**p-value**	**AUC**	**AP**	**Prec.**	**Recall**	**F1**
S	**0.69**	0.45	<0.01	**0.87** (0.74, 0.96)	0.87 (0.75, 0.96)	0.78	**0.78**	**0.78**
D	**0.71**	-0.06	0.71	0.64 (0.46, 0.79)	0.80 (0.69, 0.90)	0.80	0.57	0.67
C	**0.72**	-0.16	0.30	**0.64** (0.45, 0.79)	0.56 (0.44, 0.77)	0.56	**0.65**	**0.61**

S = Surface zone, D = Deep zone, C = Calcified zone, SC = Spearman's correlation, Prec. = Precision, F1 = F1 score

Visual inspection of the μCT images from the Test set showed that the data contains more noise compared to Cross-validation set (Figure 5). This was quantified using MSE, PSNR and SSIM. According to all metrics, Test set contained more noise (Figure 11).

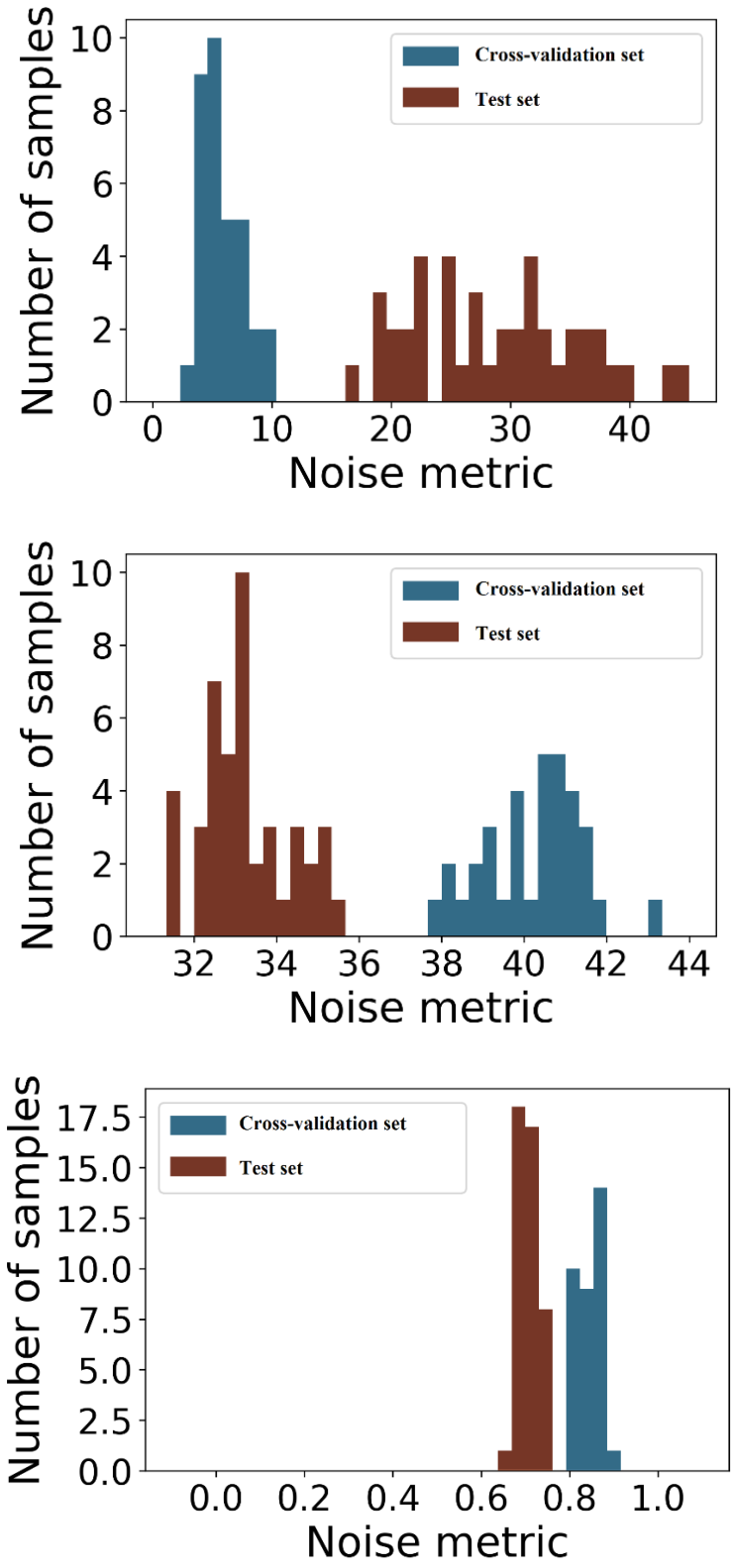

Figure 11. Results obtained from noise estimation of the datasets using MSE, PSNR and SSIM. Multiple coronal slices from each sample were compared against their median filtered counterparts. All metrics suggest that the data quality is better on the Cross-validation set. Adapted from [1].

6.4 Prototype software

During the process of developing the presented 3D grading method, the author co-developed a prototype software (Figure 12) with T. Frondelius that allows making predictions and visualizations on CEμCT imaged data. The author's responsibility was in the grading features. MRELBP is calculated using a separate MRELBP package (https://github.com/MIPT-Oulu/LocalBinaryPattern). Software utilizes the models created in the Python software to create inference on new data. Most of the software features are similar to ones used for processing of the Cross-validation set.

Additionally to the Python implementation, prototype software can be used to manually crop surface artefacts from the sample. Cropped data, as well as individual extracted VOIs can be saved for external analysis. Software is developed for Windows (C#, .NET framework, Microsoft, Visual Studio Community 2017, version 15.8.5). Its main dependencies are Activiz visualization toolkit, Accord, Microsoft Cognitive toolkit and OpenCVSharp libraries. Further details as well as all the code for the software are published on our research unit's GitHub page: https://github.com/MIPT-Oulu/3D-Histo-Grading

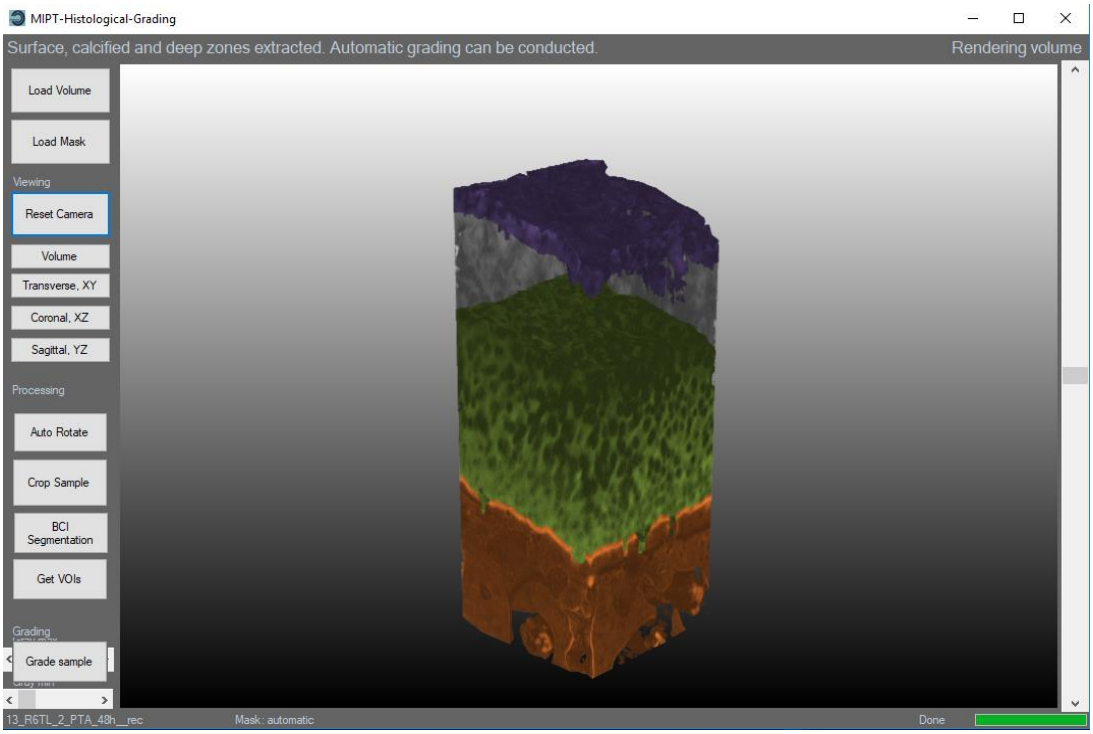

Figure 12. Example image from the prototype grading and visualization software. Software can be used to quickly make an inference on new data.

7. Discussion

In this study, the possibility to automate 3D μCT grading of osteochondral tissue was investigated for the first time. A machine learning framework was developed to assess different osteochondral zones. Models for degeneration of SZ, DZ and CZ were trained and evaluated with an independent Test set, as well as internal cross-validation to see how the approach generalizes to unseen data. Extensive statistical and quantitative analysis was used to evaluate model performance and quantify differences in the datasets.

The results obtained in this study indicate that the presented approach is best suited to detecting areas with degenerative features, rather than predicting accurate grades. This suggests that with small number of samples, models are not powerful enough to distinguish individual grades from each other, but can classify high or low level of degeneration. However, according to the experiments, the classification can even be performed on very different data acquisition protocols with high reliability.

Best performance of all zones was with the SZ models on both datasets. Performance evaluation of Cross-validation set shows better results for CZ compared to DZ model. However, inference on Test set shows opposite results. DZ models showed increase of 0.33 on AP value, while drop of 0.09 was observed for the CZ. Similar findings were obtained during parameter optimization, since most unreliability was on the CZ models for the parameter sets.

These results suggest better reliability on the DZ model. This is quite intuitive, since the DZ texture image is collapsed from a large portion of cartilage (60%) and relevant information is captured with high probability. On the contrary, CZ VOI is very thin and slight errors in the segmentation might escalate also for the final predictions. Further, the grade distributions of both DZ and CZ are very different on Test set compared to Cross-validation set (Table 1). This is certainly not the ideal case. To improve the accuracy of the DZ model, potentially a smaller proportion of the cartilage could be used to avoid including the transitional zone in the VOI. CZ model could be improved by enhancing the segmentation, or evaluating a thicker part of the calcified tissues, maybe even the full SCB volume.

Multiple optional processing and validation steps were implemented in the framework to facilitate the generalization of the presented approach to a new dataset. Some of these preprocessing steps result in lower prediction accuracy compared to what we have previously

observed. However, a method developed to only give the maximum possible performance on the training set has a high risk of overfitting, and thus being useless for new data. Additional steps include: TPE algorithm (instead of random search) on the parameter optimization, automatic artefact cropping on the texture images, normalization of the MRELBP histogram (instead of using the absolute values), PCA (that is based on explained variance or very few components) with whitening as well as averaging smaller subimages when utilizing large samples. For example, high number of PCA components (or using the 28 MRELBP features without PCA) would easily overfit to the training set, since most of the data variance is located in the first 3-5 PCA components.

Proper validation steps are important for the predictive models to ensure the model generalization. Validation can be improved with nested LOO or leaving an independent sample set outside model training for testing. High values reported using only internal cross-validation are often reduced (in worst case to levels of a random guess) when evaluated on independent testing, and such values should always be reported when possible [94]. This was also utilized in the presented study: using the additional test set, bias of the internal cross-validation could be avoided. During the study, also a third dataset was assessed for independent testing, consisting of samples from two asymptomatic cadavers. Promising results were seen for Logistic regression model on SZ. However, grade distribution was highly imbalanced and the data was excluded.

Samples from the same patients as in the Cross-validation set were also used in a previous study for cartilage surface degeneration assessment. Ylitalo et al [10] utilized PTA-stained CEμCT imaging to develop a method to quantify complex structure of the AC surface and assess obtained parameters against different manual μCT grades. They achieved high classification results for grade 0 against grade ≥ 1: surface continuity (AUC: 0.93, [0.80 0.99]), fissures (AUC: 0.94, [0.83 0.99]) and fibrillation (AUC: 0.98, [0.88 1.0]). However, surface continuity and fibrillation had only low amount of grade 0's (n = 7 against 29 in surface continuity n = 12 against 24 in fibrillation). This might overestimate the results, since low amount of negatives decreases the penalty of making false positive prediction in the ROC curve analysis.

SZ model developed in this study achieved comparable accuracy on surface continuity (AUC: 0.92 [0.80, 0.99]). Furthermore, a different split was used to balance the grade distribution (grades < 1 against ≥ 2). More extensive validation of the model was also performed by use of the PRC analysis and independent testing on another dataset. As shown in [87], PRC analysis is more informative compared to ROC curves when utilizing a biased dataset.

Conducted study has a few limitations that should be noted. Training a very reliable model could require from hundreds to thousands of samples and the presented models were trained only based on the 34 samples of the Cross-validation set, containing only TKA patients. The datasets used are very heterogeneous due to the differences in the acquisition protocol. The μCT imaging parameters were optimized for 2mm samples and doubling the diameter means that higher proportion of the X-rays attenuate on the samples, resulting in lower signal-to-noise ratio. This was assessed both visually (Figure 5), as well as quantitatively (Figure 11, Table 4). Different grade distribution was also observed on the Test set, which could be due to the low patient count. Errors in the CZ model could be due to inaccurate segmentation: k-means clustering based segmentation might not be able to capture the complex structure of the tidemark. U-Net segmentation approach was not used for the Test set, since it did not generalize with sufficient accuracy. This could be due to differences in the acquisition protocol.

In the future, depth-wise locations or thickness of the zones could be fine-tuned to find osteochondral areas that even better resemble the manually assessed grades. Furthermore, multiple evaluations from different VOIs could be combined to give an overall score for the tissue. Score like this could be used as a complementary metric to e.g. OARSI grade.

8. Conclusions

The presented thesis provides thorough introduction for using supervised learning algorithms in automating CEμCT-based volumetric grading of osteochondral tissue pathology. From the results of the study, it can be concluded that:

1) We demonstrated for the first time, that automation of the 3D CEμCT-based detection of osteochondral defects is feasible using machine learning
2) To ensure that the trained models generalize well on new data, similar data acquisition protocols should be used on all samples
3) The presented methods have potential to aid the OA researcher and pathologists by introducing objectivity in the grading process, and providing the grader a reference in cartilage assessment

All codes used in obtaining the results presented in this thesis, and in the related article will be published on our research unit's GitHub page (https://github.com/MIPT-Oulu) to allow further use of the presented methods in the OA research community.

References

[1] Rytky SJO, Tiulpin A, Frondelius T, Finnilä MAJ, Valkealahti M, Lehenkari P, et al. 3D semi-automatic osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography. 2019.

[2] Palazzo C, Nguyen C, Lefevre-Colau M, Rannou F, Poiraudeau S. Risk factors and burden of osteoarthritis. Annals of Physical and Rehabilitation Medicine 2016;59(3):134-138.

[3] Zhang W, Likhodii S, Zhang Y, Aref-Eshghi E, Harper PE, Randell E, et al. Classification of osteoarthritis phenotypes by metabolomics analysis. BMJ Open 2014;4(11):e006286.

[4] Aho O, Finnilä M, Thevenot J, Saarakkala S, Lehenkari P. Subchondral bone histology and grading in osteoarthritis. PloS one 2017;12(3):e0173726.

[5] Pritzker KP, Gay S, Jimenez SA, Ostergaard K, Pelletier JP, Revell PA, et al. Osteoarthritis cartilage histopathology: grading and staging. Osteoarthritis Cartilage 2006;14(1):13-29.

[6] Nieminen HJ, Gahunia HK, Pritzker KPH, Ylitalo T, Rieppo L, Karhula SS, et al. 3D histopathological grading of osteochondral tissue using contrast-enhanced micro-computed tomography. Osteoarthritis Cartilage 2017;25(10):1680-1689.

[7] Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. Scientific Reports 2018;8(1):1727.

[8] Tiulpin A, Klein S, Bierma-Zeinstra SMA, Thevenot J, Rahtu E, van Meurs J, et al. Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. eprint arXiv:1904.06236 2019:arXiv:1904.06236.

[9] Ashinsky BG, Bouhrara M, Coletta CE, Lehallier B, Urish KL, Lin P, et al. Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. J Orthop Res 2017;35(10):2243-2250.

[10] Ylitalo T, Finnilä MAJ, Gahunia HK, Karhula SS, Suhonen H, Valkealahti M, et al. Quantifying Complex Micro- Topography of Degenerated Articular Cartilage Surface by Contrast- Enhanced Micro- Computed Tomography and Parametric Analyses. J Orthop Res 2019;0.

[11] Maerz T, Newton MD, Matthew HWT, Baker KC. Surface roughness and thickness analysis of contrast-enhanced articular cartilage using mesh parameterization. Osteoarthritis and Cartilage 2016;24(2):290-298.

[12] Kauppinen S, Karhula SS, Thevenot J, Ylitalo T, Rieppo L, Kestilä I, et al. 3D morphometric analysis of calcified cartilage properties using micro-computed tomography. Osteoarthritis and Cartilage 2019;27(1):172-180.

[13] Kerckhofs G, Sainz J, Maréchal M, Wevers M, Van de Putte T, Geris L, et al. Contrast-Enhanced Nanofocus X-Ray Computed Tomography Allows Virtual Three-Dimensional

Histopathology and Morphometric Analysis of Osteoarthritis in Small Animal Models. Cartilage 2014;5(1):55-65.

[14] Sophia Fox AJ, Bedi A, Rodeo SA. The basic science of articular cartilage: structure, composition, and function. Sports Health 2009;1(6):461-468.

[15] Poole CA. Articular cartilage chondrons: form, function and failure. J Anat 1997 Jul;191 (Pt 1):1-13.

[16] Minns RJ, Steven FS. The collagen fibril organization in human articular cartilage. J Anat 1977;123(2):437-457.

[17] Benninghoff A. Form und Bau der Gelenkknorpel in ihren Beziehungen zur Funktion - Zweiter Teil: Der Aufbau des Gelenkknorpels in seinen Beziehungen zur Funktion. Z Zellforsch 1925;2(5):783-862.

[18] Bullough P, Goodfellow J. The significance of the fine structure of articular cartilage. J Bone Joint Surg Br 1968;50(4):852-857.

[19] Madry H, van Dijk CN, Mueller-Gerbl M. The basic science of the subchondral bone. Knee Surg Sports Traumatol Arthroscopy 2010;18(4):419-433.

[20] Alford JW, Cole BJ. Cartilage Restoration, Part 1: Basic Science, Historical Perspective, Patient Evaluation, and Treatment Options. Am J Sports Med 2005;33(2):295-306.

[21] Barrett-Jolley R, Lewis R, Fallman R, Mobasheri A. The emerging chondrocyte channelome. Frontiers in physiology 2010;1:135.

[22] Clark JM. The structure of vascular channels in the subchondral plate. J Anat 1990;171:105-115.

[23] Loeser RF, Goldring SR, Scanzello CR, Goldring MB. Osteoarthritis: A Disease of the Joint as an Organ. Arthritis Rheum 2012;64(6):1697-1707.

[24] Aspden RM, Saunders FR. Osteoarthritis as an organ disease: from the cradle to the grave. Eur Cell Mater 2019;37:74-87.

[25] Findlay DM, Kuliwaba JS. Bone-cartilage crosstalk: A conversation for understanding osteoarthritis. Bone Res 2016;4.

[26] Goldring MB, Goldring SR. Articular cartilage and subchondral bone in the pathogenesis of osteoarthritis. Ann N Y Acad Sci 2010;1192(1):230-237.

[27] Saarakkala S, Julkunen P, Kiviranta P, Mäkitalo J, Jurvelin JS, Korhonen RK. Depth-wise progression of osteoarthritis in human articular cartilage: investigation of composition, structure and biomechanics. Osteoarthritis Cartilage 2010;18(1):73-81.

[28] Alturkistani HA, Tashkandi FM, Mohammedsaleh ZM. Histological Stains: A Literature Review and Case Study. Global journal of health science 2015;8(3):72-79.

[29] Kiviranta J, Tammi M, Jurvelin J, Säämänen A-, Helminen HJ. Fixation, decalcification, and tissue processing effects on articular cartilage proteoglycans. Histochemistry 1984 /11/01;80(6):569-573.

[30] Loqman MY, Bush PG, Farquharson C, Hall AC. A cell shrinkage artefact in growth plate chondrocytes with common fixative solutions: importance of fixative osmolarity for maintaining morphology. Eur Cell Mater 2010;19:214-227.

[31] Glasson SS, Chambers MG, Van Den Berg, W. B., Little CB. The OARSI histopathology initiative – recommendations for histological assessments of osteoarthritis in the mouse. Osteoarthritis and Cartilage 2010;18:S23.

[32] Mankin HJ, Dorfman H, Lippiello L, Zarins A. Biochemical and metabolic abnormalities in articular cartilage from osteo-arthritic human hips. II. Correlation of morphology with biochemical and metabolic data. J Bone Joint Surg Am 1971 Apr;53(3):523-537.

[33] O'Driscoll SW, Keeley FW, Salter RB. Durability of regenerated articular cartilage produced by free autogenous periosteal grafts in major full-thickness defects in joint surfaces under the influence of continuous passive motion. A follow-up report at one year. J BONE JT SURG SER A 1988;70(4):595-606.

[34] Pauli C, Whiteside R, Heras FL, Nesic D, Koziol J, Grogan SP, et al. Comparison of cartilage histopathology assessment systems on human knee joints at all stages of osteoarthritis development. Osteoarthritis and Cartilage 2012;20(6):476-485.

[35] Waldstein W, Perino G, Gilbert SL, Maher SA, Windhager R, Boettner F. OARSI osteoarthritis cartilage histopathology assessment system: A biomechanical evaluation in the human knee. J Orthop Res 2016;34(1):135-140.

[36] Kemerink M, Dierichs TJ, Dierichs J, Huynen HJM, Wildberger JE, van Engelshoven, Jos M. A., et al. Characteristics of a First-Generation X-Ray System. Radiology 2011;259(2):534-539.

[37] Medical X-ray imaging with scattered photons. Proceedings of SPIE - The International Society for Optical Engineering; 2002.

[38] Lusic H, Grinstaff MW. X-ray-computed tomography contrast agents. Chem Rev 2013;113(3):1641-1666.

[39] Schambach SJ, Bag S, Schilling L, Groden C, Brockmann MA. Application of micro-CT in small animal imaging. Methods 2010;50(1):2-13.

[40] Feldkamp LA, Davis LC, Kress JW. Practical cone-beam algorithm. J Opt Soc Am A 1984;1(6):612-619.

[41] Kalender WA. Technical foundations of spiral CT. Seminars in Ultrasound, CT and MRI 1994;15(2):81-89.

[42] Henwood S. Clinical CT : Techniques and Practice. London: Cambridge University Press; 1999.

[43] Willmott P. An introduction to synchrotron radiation : techniques and applications. Chichester, West Sussex, UK: John Wiley; 2011. p. 303-311.

[44] Yang K, Kwan ALC, Miller DF, Boone JM. A geometric calibration method for cone beam CT systems. Med Phys 2006;33(6):1695-1706.

[45] Beister M, Kolditz D, Kalender WA. Iterative reconstruction methods in X-ray CT. Phys Med 2012;28(2):94-108.

[46] Ketola JH, Karhula SS, Finnilä MAJ, Korhonen RK, Herzog W, Siltanen S, et al. Iterative and discrete reconstruction in the evaluation of the rabbit model of osteoarthritis. Scientific Reports 2018;8(1):12051.

[47] Hounsfield GN. Computerized transverse axial scanning (tomography): I. Description of system. Br J Radiol 1973;46(552):1016-1022.

[48] Silvast TS, Jurvelin JS, Aula AS, Lammi MJ, Töyräs J. Contrast Agent-Enhanced Computed Tomography of Articular Cartilage: Association with Tissue Composition and Properties. Acta Radiologica 2009;50(1):78-85.

[49] Karhula SS, Finnilä MA, Freedman JD, Kauppinen S, Valkealahti M, Lehenkari P, et al. Micro-Scale Distribution of CA4+ in Ex vivo Human Articular Cartilage Detected with Contrast-Enhanced Micro-Computed Tomography Imaging. Front Phys 2017;5.

[50] Metscher BD. MicroCT for developmental biology: a versatile tool for high-contrast 3D imaging at histological resolutions. Dev Dyn 2009;238(3):632-640.

[51] Nieminen HJ, Ylitalo T, Karhula S, Suuronen J, Kauppinen S, Serimaa R, et al. Determining collagen distribution in articular cartilage using contrast-enhanced micro-computed tomography. Osteoarthritis Cartilage 2015;23(9):1613-1621.

[52] Karhula SS, Finnilä MA, Lammi MJ, Ylärinne JH, Kauppinen S, Rieppo L, et al. Effects of Articular Cartilage Constituents on Phosphotungstic Acid Enhanced Micro-Computed Tomography. PLoS ONE 2017;12(1):e0171075.

[53] Ashton JR, Clark DP, Moding EJ, Ghaghada K, Kirsch DG, West JL, et al. Dual-energy micro-CT functional imaging of primary lung cancer in mice using gold and iodine nanoparticle contrast agents: a validation study. PLoS One 2014 Feb 10;9(2):e88129.

[54] Ashton JR, West JL, Badea CT. In vivo small animal micro-CT using nanoparticle contrast agents. Front Pharmacol 2015;6:256.

[55] Zhu Y, Manske SL, Boyd SK. Cartilage imaging of a rabbit knee using dual-energy X-ray microscopy and 1.0 T and 9.4 T magnetic resonance imaging. Journal of Orthopaedic Translation 2015 October 1,;3(4):212-218.

[56] Bailly L, Cochereau T, Orgéas L, Henrich Bernardoni N, Rolland du Roscoat S, McLeer-Florin A, et al. 3D multiscale imaging of human vocal folds using synchrotron X-ray microtomography in phase retrieval mode. Scientific Reports 2018 Dec;8(1):1-20.

[57] Bidola P, Morgan K, Willner M, Fehringer A, Allner S, Prade F, et al. Application of sensitive, high-resolution imaging at a commercial lab-based X-ray micro-CT system using propagation-based phase retrieval. Journal of Microscopy 2017 -05-01;266(2):211-220.

[58] Paganin D, Mayo SC, Gureyev TE, Miller PR, Wilkins SW. Simultaneous phase and amplitude extraction from a single defocused image of a homogeneous object. J Microsc 2002;206(1):33-40.

[59] Yu B, Weber L, Pacureanu A, Langer M, Olivier C, Cloetens P, et al. Evaluation of phase retrieval approaches in magnified X-ray phase nano computerized tomography applied to bone tissue. Opt Express, OE 2018 /04/30;26(9):11110-11124.

[60] Shankle WR, Mania S, Dick MB, Pazzani MJ. Simple models for estimating dementia severity using machine learning. Stud Health Technol Informatics 1998;52.

[61] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informatics 2006;2:59-77.

[62] Dürr O, Sick B. Single-cell phenotype classification using deep convolutional neural networks. J Biomol Screen 2016;21(9):998-1003.

[63] Lötsch J, Ultsch A. A machine-learned knowledge discovery method for associating complex phenotypes with complex genotypes. Application to pain. J Biomed Informatics 2013;46(5):921-928.

[64] Chen C-. A note on sequential decision approach to pattern recognition and machine learning. Information and Control 1966;9(6):549-562.

[65] Theodoridis S. Machine Learning : A Bayesian and Optimization Perspective. London: Academic Press; 2015.

[66] Kulkarni S, Harman G, Wiley InterScience (Online service). An Elementary Introduction to Statistical Learning Theory. Hoboken, N.J.: Wiley; 2011.

[67] Vapnik VN. Statistical learning theory. New York: Wiley; 1998. p. 736 sivua.

[68] Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970;12(1):55-67.

[69] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software 2010;33(1):1-22.

[70] I. Naseem, R. Togneri, M. Bennamoun. Linear Regression for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2010;32(11):2106-2112.

[71] Hastie T, Friedman J, Tibshirani R. The elements of statistical learning : data mining, inference, and prediction. 2nd ed ed. New York: Springer; 2009. p. 745 sivua.

[72] Malley JD, Pajevic S, Malley KG. Statistical Learning for Biomedical Data. Cambridge: Cambridge University Press; 2011.

[73] Luo Y, Wu C, Zhang Y. Facial expression feature extraction using hybrid PCA and LBP. The Journal of China Universities of Posts and Telecommunications 2013;20(2):120-124.

[74] Luo Y, Wu C, Zhang Y. Facial expression recognition based on fusion feature of PCA and LBP with SVM. Optik - International Journal for Light and Electron Optics 2013;124(17):2767-2770.

[75] Zhou SK, Zhao W, Tang X, Gong S, editors. Fusing Gabor and LBP Feature Sets for Kernel-Based Face Recognition. Analysis and Modeling of Faces and Gestures Berlin, Heidelberg: Springer Berlin Heidelberg; 2007.

[76] T. Ojala, M. Pietikainen, T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002;24(7):971-987.

[77] Finnilä MAJ, Thevenot J, Aho O-, Tiitu V, Rautiainen J, Kauppinen S, et al. Association between subchondral bone structure and osteoarthritis histopathological grade. J Orthop Res 2017;35(4):785-792.

[78] L. Liu, S. Lao, P. W. Fieguth, Y. Guo, X. Wang, M. Pietikäinen. Median Robust Extended Local Binary Pattern for Texture Classification. IEEE Transactions on Image Processing 2016;25(3):1368-1381.

[79] Liu L, Fieguth P, Guo Y, Wang X, Pietikäinen M. Local binary features for texture classification: Taxonomy and experimental study. Pattern Recognition 2017;62:135-160.

[80] Tihonov AN. On the solution of ill-posed problems and the method of regularization. Dokl Akad Nauk SSSR 1963;151:501-504.

[81] Feature selection, L1 vs. L2 regularization, and rotational invariance. Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004; 2004.

[82] Santosa F, Symes W. Linear Inversion of Band-Limited Reflection Seismograms. SIAM J Sci and Stat Comput 1986;7(4):1307-1330.

[83] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society.Series B (Methodological) 1996;58(1):267-288.

[84] A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. : Morgan Kaufmann; 1995.

[85] Ivanescu AE, Li P, George B, Brown AW, Keith SW, Raju D, et al. The importance of prediction model validation and assessment in obesity and nutrition research. International journal of obesity (2005) 2016;40(6):887-894.

[86] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 1997;30(7):1145-1159.

[87] Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 2015 Mar 4,;10(3):e0118432.

[88] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci 2015;9351.

[89] LAPACK: A Portable Linear Algebra Library for High-performance Computers. Proceedings of the 1990 ACM/IEEE Conference on Supercomputing Los Alamitos, CA, USA: IEEE Computer Society Press; 1990.

[90] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281-305.

[91] Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011; 2011.

[92] Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. : JMLR; 2013.

[93] Semi-supervised learning for semantic relation classification using stratified sampling strategy. EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009; 2009.

[94] T. Upadhaya, M. Vallières, A. Chatterjee, F. Lucia, P. A. Bonaffini, I. Masson, et al. Comparison of Radiomics Models Built Through Machine Learning in a Multicentric Context With Independent Testing: Identical Data, Similar Algorithms, Different Methodologies. IEEE Transactions on Radiation and Plasma Medical Sciences 2019;3(2):192-200.